Analysis of Metabolic Alterations during Colorectal Cancer Development

Dissertation

zur

Erlangung der naturwissenschaftlichen Doktorwürde

(Dr. sc. nat.)

vorgelegt der

Mathematisch-naturwissenschaftlichen Fakultät

der

Universität Zürich

von

David Jonathan Fischer

aus

Deutschland

Promotionskomitee

Prof. Dr. Josef Jiricny (Vorsitz der Dissertation)

Dr. sc. nat. ETH Endre Laczko (Leiter der Dissertation)

PD Dr. Giancarlo Marra

Prof. Dr. Reinhard Furrer

Dr. Christian Frezza (Externer Gutachter)

Zürich 2014

Zusammenfassung

Schon seit Beginn des zwanzigsten Jahrhunderts ist bekannt, dass Tumoren im Vergleich zu gesundem Gewebe einen veränderten Stoffwechsel haben. Allerdings ist erst während den letzten Jahrzenten bekannt geworden, dass Krebszellen ihren Stoffwechsel als Folge von Anhäufung genetischer Mutationen umprogrammieren können. Entstehung von Darmkrebs ist ein klassisches Beispiel für die Tumorigenese und beschreibt die Entwicklung normaler Schleimhaut über verschiedene Stadien gutartiger Tumoren (Adenomen) hin zu bösartigem Darmkarzinom. Viele der genetischen Veränderungen, die zur Entwicklung dieser einzelnen Stadien führen sind bekannt und für einige der am häufigsten mutierten Gene in Darmkrebs, *MYC*, *TP53* or *KRAS,* konnte eine Verbindung zu einer Stoffwechseländerung in Krebs nachgewiesen werden. Allerdings ist unklar, wie diese genetischen Faktoren zur Veränderung des Stoffwechsels während der Darmkrebs-entstehung beitragen. Ziel dieser Dissertation ist es, die metabolischen Veränderungen in gutartigen Darmtumoren und Darmkrebs anhand von Metabolomics- und Transcriptomics Analysen explorativ zu untersuchen. Da die Grösse der verwendeten Darmbiopsien sehr klein ist, beschreiben die ersten Kapitel die Entwicklung neuartiger Methoden zur umfassenden Analyse des Metaboloms und Lipidoms aus kleinen Probenmengen. Es wurden zwei sich ergänzende Methoden entwickelt, die auf Kapillarfluss-Flüssigchromatographie, gekoppelt mit Massenspektrometrie (capLCMS) basieren. Beide Methoden konnten eine sensitive Quantifizierung der meisten getesteten Reinsubstanzen im femtomolaren Bereich ermöglichen. Angewendet auf Metabolitenextrakte von Darmgewebe konnten wir zeigen, dass 50% der detektierten Massen in der KEGG Metabolitendatenbank auffindbar waren, ausserdem 700 verschiedene Lipide. Relative Standard-abweichungen der quantifizierten Metaboliten lagen im Bereich zwischen 15 und 20%, was einer reproduzierbaren Methode entspricht und es ermöglicht, die Metabolomunterschiede zwischen Darmkrebs und Normalgewebe zu untersuchen. Desweiteren wird die Entwicklung einer neuen Technik zur Auswertung von LCMS Rohdaten beschrieben. Diese Technik hat das Ziel, Metaboliten mit geringem Vorkommen besser zu detektieren. Anstelle jede einzelne Rohdatei einer LCMS-Analyse individuell zu bearbeiten, ist die Idee

dieser Methode, alle gewonnenen LCMS-Daten zusammenzuführen und Metaboliten auf dem so entstandenen kombinierten Datensatz zu identifizieren. Die entwickelte Methode wurde mit einer bestehenden Software, XCMS, verglichen. In diesem Vergleich zeigte sich, dass unsere Methode bis zu zehnmal mehr Metaboliten detektieren kann als XCMS. Darüber hinaus waren die Intensitäten der Metabolitensignale, die nur durch unsere Methode identifizierbar waren, zehnmal geringer. Diese Beobachtungen zeigen, dass der Ansatz, LCMS Daten zusammenzuführen tatsächlich dazu führen kann, Metaboliten mit geringer Menge besser zu detektieren. Anschliessend wurden die entwickelten Metabolomics und Lipidomics Methoden, zusammen mit bereits vorhandenen Transcriptomics Daten dazu verwendet, die Stoffwechseländerungen in Darmkrebs und Darmkrebsvorstufen zu untersuchen. Interessanterweise stellte sich heraus, dass die meisten Änderungen in der Expression von Enzymgenen und die Änderung der Metabolitenkonzentration zwischen Adenom und Darmkrebs sehr ähnlich waren, was erklären würde dass die Änderung im Stoffwechsel schon in frühen Darmkrebsstadien stattfindet. Zu den beobachteten Stoffwechselwegen zählen Glykolyse, Prolin- Serin und Nukleotid-biosynthese, sowie Glutaminstoffwechsel. Die meisten Enzymgene, die in diesen Stoffwechselwegen involviert waren, korrelierten mit der Expression des *MYC* Gens, ausserdem mit Zielgenen von c-Myc selbst, was darauf hinweist dass diese Gene durch den wnt-Signalweg kontrolliert sind. Andere beobachtete Änderungen in der Genexpression oder in den Metabolitenkonzentrationen waren nur in Darmkrebs zu beobachten, aber nicht in den Adenomen. Hierzu zählt die Hochregulierung der solute carrier Gene *SLC7A5, SLC2A3* or *SLC2A1,* welche für die Aufnahme von Glukose und neutraler Aminosäuren verantwortlich sind. Ausserdem wurde eine Hochregulierung von Genen für die Synthese sehr langer, mehrfach ungesättigter Fettsäuren (VLC-PUFA) gezeigt. Diese Beobachtung wurde dadurch untermauert, dass die Anzahl der meisten Lipidklassen mit konjugierten VLC-PUFA in Darmkrebs zunahm, während Lipide mit kurzen und gesättigten Fettsäuren abnahm. Zusammen verweisen diese Ergebnisse auf eine verstärkte Biosynthese mehrfach ungesättigter Fettsäuren in Darmkrebszellen.

Summary

It was already shown at the beginning of the twentieth century that tumors have a different metabolic phenotype compared to normal cells. However only during the last decade it became evident that cancer cells undergo a metabolic reprogramming because of accumulations of genetic mutations and subsequent alterations in molecular signalling pathways which ultimately lead to abnormal cell proliferation. Development of colorectal cancer is a classical example for tumorigenesis and describes the transition of normal mucosa to a malignant colorectal carcinoma via different stages of premalignant lesions (adenomas). Many of the genetic alterations which lead to each transition step are known and some of the most commonly altered genes in colorectal cancer, *MYC*, *TP53* or *KRAS*, have been linked to the reprogramming of cancer metabolism. However, it is unclear how these factors contribute to metabolic alterations during colorectal cancer development.

This thesis aims for an exploratory analysis of metabolic changes in human colorectal adenomas and cancers by applying untargeted metabolomics and transcriptomics techniques. Because of limited biological sample sizes, the first two chapters of the thesis describe the development of novel technologies for a comprehensive analysis of the metabolome and lipidome from a small amount of colorectal biopsies. Two complementary capillary liquid chromatography mass spectrometry (capLCMS) methods were developed, both providing a sensitive quantification of most tested standard metabolites in the femtomolar range. Both methods cover about 50% of the masses annotated in the KEGG metabolite database and about 700 lipids. Relative standard deviations lie in the range between 15 and 20%, which makes the methods reproducible and applicable for the comparison between tumors and normal mucosa samples.

We then describe the development of a novel LCMS raw data processing technique, which aims to an improved quantification of low abundant metabolites. The basic idea of the proposed algorithm is to merge multiple LCMS runs together and to detect metabolites on this merged dataset instead of in each single file. We compared our algorithm with existing software. We

demonstrate that our approach yields up to ten times more metabolite features than the open source software tool XCMS. Moreover, metabolites detected uniquely by the new method, were ten times lower in intensity as the lowest abundance metabolites detected by XCMS. This indicates that the new algorithm is indeed more sensitive towards low abundant metabolites.

Finally, we applied our lipidomics and metabolomics techniques to investigate the metabolic alteration of colorectal tumors, in combination with previously acquired transcriptomics data. Most interestingly, many alterations in gene expression of enzymes and in metabolite abundance were similar in adenomas and cancers, indicating that reprogramming of metabolism occurs early in colorectal tumorigenesis. These alterations included metabolic pathways involved in glycolysis, nucleotide biosynthesis, and glutamine, proline and serine metabolism. Most of the enzyme genes expressed in these pathways correlated with the expression of *MYC* and its target genes, which indicates that theses enzymes are controlled by the Wnt pathway. Other alterations in gene expression or metabolite abundance were observable in cancer but not in adenomas, such as the upregulation of solute carrier genes *SLC7A5, SLC2A3* or *SLC2A1* whose products are involved in the uptake of glucose and neutral amino acids. Also, the expression of genes encoding proteins that are involved in the synthesis of very long, polyunsaturated fatty acids (VLC-PUFA) were more increased in cancer than in adenomas. Complementing this observation, most of the lipids containing VLC-PUFA were increased in cancers, while lipids with short and saturated fatty acyls were decreased, indicating that cancer cells have an increased biosynthetic activity for VLC-PUFA.

Table of Contents

Introduction

Colorectal Cancer

Colon and rectal cancer (colorectal cancer, CRC) is worldwide the third most commonly diagnosed cancer (Jemal et al., 2011) and CRC related death rates tend to decline in developed countries, where incidence is more common (Siegel et al., 2013). These declines in mortalities reflect an improvement in early diagnostics and treatment of this disease during the last 40 years (Cress et al., 2006). However, CRC still remains the third most frequent cancer fatality in the USA (Siegel et al., 2013).

CRC is believed to arise from premalignant neoplastic lesions (adenomas) and the progress which leads to transformation of normal mucosa to malignant tumors was extensively studied in the past (Fearon and Vogelstein, 1990). Hundreds of somatic mutations are acquired during tumor development, among which only a small subset of "driver mutations" are suspected to be responsible for the progression of cancer (Stratton et al., 2009). There are three main driver gene mutations involved in a prominent fraction of CRC: *APC*, *KRAS* and *TP53* (Fearon, 2011). These main driver genes in CRC are already known for several years and even after most recent genome screenings there are no other mutation frequencies as high as in *KRAS* (45.1%), *TP53* (58.6%) and *APC* (81.9%) (Kandoth et al., 2013). CRC is typically divided into two subclasses: About 85% of CRC cases are considered as chromosomally unstable as they contain frequent allelic losses. *APC*, *KRAS* and *TP53* somatic mutations are typically found in patients having these tumor forms. The other 15% of the cases typically show 10-fold more overall mutations including microsatellite instability (MSI) due to a defect of the DNA mismatch repair. These hypermutated tumors often contain mutations or epigenetic silencing in the mismatch repair genes *MLH1*, *PMS2*, *MSH2*, and *MSH6* (Marra and Schar, 1999; Truninger et al., 2005; Muzny et al., 2012). It is estimated that 20% CRC cases have a hereditary background, where familial adenomatous polyposis (FAP) and hereditary nonpolyposis colorectal cancer (HNPCC) are the major mendelian forms (Lynch and la Chapelle, 2003). The most common form HNPCC, also termed Lynch syndrome, is typically characterized by MSI

positive tumors and by germline mutations of one of the mismatch repair genes. FAP is an autosomal dominantly inherited disease mostly caused by germline mutations in *APC* (>90%) and patients develop hundreds of adenomas early in their life. In sporadic CRC, novel somatic mutations lead first to the development of adenomas and further independent mutations lead finally to cancer, a progress which is estimated to take around 20 years and CRC is therefore typically diagnosed only in elder people (Jones et al., 2008; Vogelstein et al., 2013). The genetic predisposition of FAP and HNPCC however accelerates CRC development either at the formation of adenomas (FAP) or at the transition from adenomas to carcinomas (HNPCC).

The inactivation of *APC* or other Wnt-signalling genes is considered as the first key event in colorectal tumorigenesis (Clevers and Nusse, 2012). Stem cells within the intestinal crypts are normally responsible for the renewal of the mucosal epithelium. Activation of the Wnt-pathway drives the transformation of these cells and leads to the formation of neoplasia (Barker et al., 2008). The most common form of such colorectal neoplasia are adenomatous polypoid lesions (or polyps) and are characterized by protruding growth into the gut lumen, either sessile or with a stalk (Cattaneo et al., 2010). Other lesion forms do not protrude into the gut lumen and are either slightly elevated, or flat or depressed. Despite the huge progress in molecular biology and clinical management of colorectal cancers, the removal of precancerous lesions during colonoscopy is considered as the most efficient and reliable tool to reduce CRC incidence and mortality (Winawer et al., 1993; Lena, 2013).

Cancer Metabolism

Today it is widely acknowledged that cancer is a genetic disease and therefore most tumors are believed to develop from alterations in the genome of dedicated tumor cells, for example by activation or deactivation of signalling pathway genes (Vogelstein and Kinzler, 2004). From a historical perspective however, the first molecular alterations of cancer were discovered on the level of metabolism (Warburg, 1926). In the 1920s, Otto Warburg showed that tumors metabolize glucose via glycolysis instead of oxidative respiration (Koppenol et al., 2011). The resulting metabolic phenotype was the

production of lactic acid, even if enough oxygen was supplied. This phenomenon that cancer cells produce lactic acid from glucose is now widely appreciated as Warburg effect. For a long time, Warburg believed that the origin of cancer cells are explained by an impairment of respiration, a controversial hypothesis which was hotly debated until the 1970s (Warburg, 1956; Weinhouse et al., 1956; Weinhouse, 1976). After Warburg's death in 1970 the field of cancer research was greatly influenced by the discovery of oncogenes and tumor suppressor genes (Weinberg, 1994). This basic concept of cancer cell proliferation by gene activation (oncogene) or by inhibiting activity of gene product (tumor suppressor gene) is until today a basic dogma of tumorigenesis (Vogelstein and Kinzler, 2004). During these years the interest in cancer metabolism decreased due to the popularity and the importance of molecular genetics in cancer (Figure i). However, during the last decade cancer metabolism experienced a renaissance within the cancer research community (Figure i), and changes in energy metabolism has been acknowledged as one of the prominent hallmarks of cancer (Hanahan and Weinberg, 2011). This recent gain of interest is explained by the discovery of tight associations between genetic alterations and metabolism in cancer cells. Among these genetic-metabolic associations we can distinguish between two kinds of observations: 1) Cancer-associated mutations of enzyme genes and 2) Transcriptional or post transcriptional regulation of enzymatic genes via oncoproteins and tumor suppressors. Interestingly, most of the so far discovered cases of direct mutations in enzyme genes represent mitochondrial genes involved in TCA cycle (Koppenol et al., 2011). Mutations in fumarate hydratase (*FH*), succinate dehydrogenase (*SDH*) and isocitrate dehydrogenase (*IDH1*) are an indication that Warburg's hypothesis of defective mitochondria as cause of cancer may hold true for a few cancer types (Baysal et al., 2000; Tomlinson et al., 2002; Parsons et al., 2008). In many other cases however, metabolic reprogramming of cancer cells is mediated by oncogenes and tumor suppressor genes (Ward and Thompson, 2012). Glycolysis is one target of metabolic reprogramming, and many glycolytic genes are upregulated in cancer (Kroemer and Pouyssegur, 2008). Upregulated glycolytic activity, and especially the increase in lactate dehydrogenase confirm the Warburg effect. In addition to glycolysis, cancer cells rely on biosynthetic processes in order to provide building blocks for

DNA, lipid and protein biomass (DeBerardinis et al., 2008). Many major oncogenes and tumor suppressor genes such as p53, HIF1, AKT, RAS, VHL, c-MYC are directly linked to reprogramming of metabolism in cancer (Koppenol et al., 2011). In addition to glucose, glutamine plays a major role as means of participating in bioenergetic and biosynthethic reactions in cancer (DeBerardinis and Cheng, 2009). Glutaminolysis and related metabolic processes such as nucleotide biosynthesis are mainly regulated by c-Myc (Wise et al., 2008). Some very recent findings report the rediscovery of metabolic pathways, such as those of glycine and proline to be involved in metabolic reprogramming of cancer cells (Liu et al., 2012; Locasale, 2013).

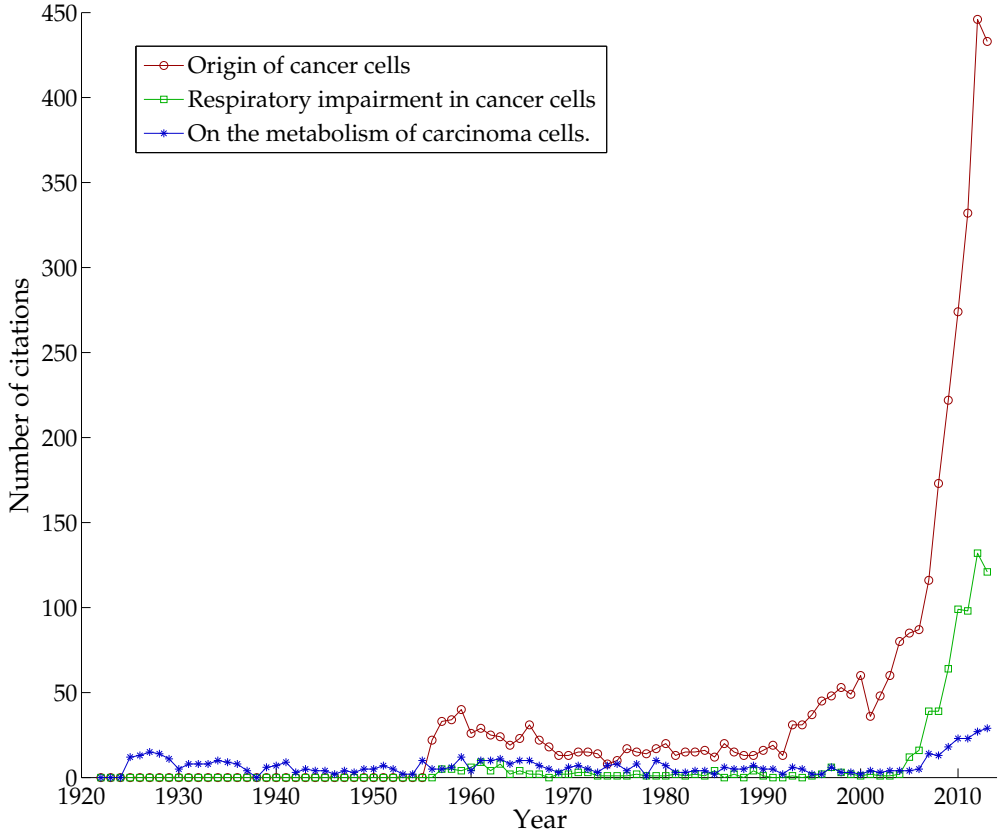


Figure i: Interest in cancer metabolism research between 1920 and 2014 indicated by the three most highly cited publications of Otto Warburg in this topic. Citation peaks are between 1920 and 1930, between 1955 and 1965 and from 2005 until today. Origin of cancer cells (Warburg, 1956), Respiratory impairment in cancer cells (Weinhouse et al., 1956), on the metabolism of carcinoma cells (Warburg, 1925). Data was retrieved from https://www.webofknowledge.com.

Investigation of metabolism using mass spectrometry based metabolomics

Metabolites are small molecules with a molecular weight typically below 2000 Da. As a product of enzymatic activity they are the ultimate products of a biological system (Fiehn, 2002). Metabolites are chemically very heterogenous and include different classes such as amino acids, lipids, nucleotides or central carbon metabolites (Milne et al., 2013). In addition to these very common classes, there exists a vast amount of secondary metabolites, for example the many diverse natural products of plants (Roepenack-Lahaye et al., 2004). Environmental factors such as chemicals, drugs or microbes additionally contribute to the overall composition of metabolites within an organism. The total sum of all these endogenous and exogenous compounds is called the metabolome. It it estimated that the human metabolome consists of about 100000 small molecules and the current version of the human metabolome database (HMDB) lists alone 29332 endogenous metabolites (Milne et al., 2013; Wishart et al., 2013). Technical advances in the fields of mass spectrometry (MS) allow us nowadays to detect and quantify thousands of metabolites, enabling us to analyse a broad range of the metabolome (Patti et al., 2012). In addition to genomics, transcriptomics and proteomics, metabolomics provides us therefore an important tool for systems biology and functional genomics (Hollywood et al., 2006).

In contrast to genomics and transcriptomics, MS based metabolomics is not capable to cover most of its individual parts, i.e. each metabolite, with a single technological setup. While novel mass spectrometers with high acquisition rates already allow for comprehensive analysis of the proteome (Hebert et al., 2014), the heterogenous nature of metabolites hamper the development of an all-encompassing method. Metabolomics is therefore typically divided into a 1) Targeted or 2) Untargeted analysis of metabolites (Patti et al., 2012). Targeted metabolomics generally aims for the quantification of a small subset of metabolites, often employing a triple quadrupole mass spectrometer (Lu et al., 2008). These instruments are then used in multiple reaction monitoring mode (MRM), where MS parameters have to be individually optimized for the metabolites of interest before the real metabolomics experiment can be performed: For each metabolite, a pure standard compound needs to be

obtained in order to optimize the following parameters: 1) A precursor ion for the isolation of the most abundant mass of each metabolite, 2) the most abundant fragment ion which is optimally highly specific for each metabolite and 3) the optimal collision energy which is needed to gain optimal yield of the fragment ion (Kitteringham et al., 2009). The advantages of this technique are a high sensitivity, good reproducibility and a broad dynamic range. Very often this method is also highly specific for each metabolite, leading to an unambiguous detection. Typically, the focus of a targeted metabolomics method is to specifically quantify a subset class of metabolites, such as central carbon metabolites (Luo et al., 2007).

Untargeted metabolomics, or metabolic profiling, has the general aim to quantify as much metabolites as possible in a biological sample (Patti et al., 2012). NMR and mass spectrometry are the most frequent analytical techniques for this approach (Dunn et al., 2005). However with modern mass spectrometers, thousands of mass peaks can be monitored with a high mass accuracy and a good sensitivity, making MS and especially liquid chromatography mass spectrometry (LCMS) the method of choice for untargeted metabolomics (Forcisi et al., 2013). The big advantage of LCMS based untargeted metabolomics is therefore the possibility to screen a large amount of metabolites, however the major bottleneck of this technique is data interpretation (Patti et al., 2013). Typically, high mass accuracy full scan LCMS data are highly complex and file sizes can approach gigabyte dimensions. In order to interpret such data, a first step is to generate data tables from the raw data, where each identified signal is annotated by its information on the mass to charge (m/z), retention time (RT) and intensity dimension (Sugimoto et al., 2012). As a next step, accurate masses from the data table are searched in metabolite databases such as the human metabolome database (Wishart et al., 2007), METLIN (Tautenhahn et al., 2012) or KEGG (Ogata et al., 1999), in order to putatively annotate possible metabolite hits.

Untargeted metabolomics made huge progress during the last decade and the modern techniques allow for a comprehensive analysis of the metabolome. The nowadays challenges in untargeted metabolomics are to improve the dynamic range and sensitivity of the analytical methods in order to cover a

broad range of metabolite classes from a low amount of biological material (Koek et al., 2010; Ivanisevic et al., 2013). One big success for the nowadays comprehensive and sensitive proteomics techniques was the development of LCMS techniques based on nanoflow and nanoelectrospray (Shen et al., 2002). Very low flow rates allow for a high chromatographic separation and very good sensitivity. Despite these advantages, nanoLCMS applications are very uncommon in metabolomics research. A few metabolomics applications using nanoLCMS demonstrate a sensitivity range which is comparable to targeted approaches or even better (Myint et al., 2009a; 2009b; Kiefer et al., 2010). However, normal flow LCMS is still the dominating analytical method in both untargeted and targeted metabolomics.

Aim of this thesis

As mentioned above, the tumor suppressor genes *APC* and *TP53*, as well as the oncogone *KRAS* play an important role for the transition from normal mucosa to adenoma and cancer (Fearon and Vogelstein, 1990). As oncogenes and tumor suppressors were reported to be involved in the reprogramming of metabolism in cancer cells, we hypothesized that metabolic alterations might be different in each stage of tumorigenesis. *TP53* was shown to be involved in glycolysis and pentose phosphate metabolism and more recently also in glutamine metabolism (Bensaad et al., 2006; Hu et al., 2010; Suzuki et al., 2010). Also more recently it was shown that *KRAS* stimulates glucose uptake and pentose phosphate biosynthesis (Ying et al., 2012). Most interestingly, as a consequence of wnt pathway deregulation in the development of premalignant colorectal tumors, the transcription factor c-Myc is overexpressed in basically all early colorectal lesions (Sansom et al., 2007; Wilkins and Sansom, 2008). Given the fact that c-Myc plays a role in many different metabolic processes (Li and Simon, 2013), we hypothesized that reprogramming of metabolism is already an early event in tumorigenesis. During the last years, a large amount of gene expression data from colorectal tumor biopsies, including adenomas and cancer were acquired (Cattaneo et al., 2010; Maglietta et al., 2012). Guided by these valuable data and motivated by the recent advances in cancer metabolism, the presented work aims for an exploratory analysis of the metabolic changes during colorectal cancer

development by a combined analysis of the metabolome and transcriptome of human colorectal tumors.

A major technical problem for the analysis of colorectal biopsies is that only a limited amount of sample is available, especially for adenomatous lesions (Sabates Bellver et al., 2007). The premise for the analysis of a colorectal tissue metabolome was therefore to establish analytical methods for the analysis of less than 5 mg biopsies. This thesis is divided into three parts: The development of two different technical strategies for the improvement of sensitivity in mass spectrometry based metabolomics and the investigation of metabolic changes of colorectal cancer development:

Chapter 1:

Decreased column diameter in LCMS analytics was described to improve the ionization, ion suppression effects and overall sensitivity compared to normal flow LCMS (Shen et al., 2002). In this chapter the development of two complementary methods based on capillary LCMS for a comprehensive analysis of the lipidome and metabolome are presented and discussed.

Chapter 2:

Reliable and sensitive computational methods are essential for the analysis of untargeted metabolomics LCMS data. In this chapter we will discuss the development of a novel algorithm with an increased sensitivity towards low abundant metabolites.

Chapter 3:

In this chapter we will combine transcriptomics and metabolomics data of colorectal biopsies in order to investigate the metabolic changes during colorectal cancer development.

1 Fast and sensitive metabolomics and lipidomics of low abundant biological material using capillary scale LCMS

David Jonathan Fischer[1], Giancarlo Marra[2] and Endre Laczko[1]

[1]Functional Genomics Center Zürich and [2]Institute of Molecular Cancer Research, University of Zürich, 8057 Zürich, Switzerland

D.J. Fischer contributed to the design of the experiments, analysed the data and wrote the manuscript. G. Marra and E. Laczko supervised the study and contributed to the experimental design

1.1 Abstract

Metabolomics is considered as a complementary method to proteomics and transcriptomics. The comprehensive, efficient identification and quantification of the metabolites and lipids in any given biological sample is still a major challenge within the field, especially when samples of limited volume are to be analyzed. Here we present a complementary set of untargeted metabolomics methods which are based on capillary scale ultra-performance liquid chromatography mass spectrometry (capLCMS) for the comprehensive analysis of metabolites ranging from very polar (sugars) to very non-polar (lipids) compounds. Two chromatographic setups were chosen, one based on hydrophilic interaction chromatography (HILIC) and the other on reversed phase chromatography (RP). Limits of detection for both methods were shown to be in the low picomolar to femtomolar range for most of a selection of 190 reference compounds, covering all major metabolite classes. We demonstrate a broad metabolome and lipidome coverage by applying both methods to colorectal tissue extracts from biopsies with less than 5 mg fresh weight. About 50% of the KEGG metabolome database and 688 lipids could be annotated by accurate mass and with good relative standard deviation (RSD) values between 15 and 20%. A total run time of only 35 minutes for both methods allows for a high throughput analysis.

1.2 Introduction

Metabolomics aims for a comprehensive quantification of small molecules in a biological system (Baker, 2011). With technological advances during the last years, liquid chromatography coupled to mass spectrometry (LCMS) became a powerful tool to measure a broad range of metabolites (Patti et al., 2012). In contrast to the analysis of typically hydrophobic peptides in shotgun-proteomics, small molecules are chemically very heterogeneous and can range from very polar (e.g. sugars) to very non-polar (lipids) compounds. Therefore, quantification of the complete metabolome is impossible to achieve in a single analytical setup. RP chromatography is a very popular and powerful analytical method but is typically restricted to the analysis of lipids and other hydrophobic molecules (Lu et al., 2008). For the analysis of polar metabolomes, hydrophilic stationary phases (Bajad et al., 2006; Pesek et al.,

2008) or reversed phase chromatography which include ion pairing reagents as a polar anchor (Coulier et al., 2006) were described.

A further analytical challenge is sensitivity. For many biological applications, metabolite extracts originate from a small amount of cells or small tissues, and therefore it is desirable that LCMS systems are able to analyse low amounts of metabolites. One strategy to gain sensitivity is the reduction of the chromatographic column dimensions, i.e. column diameter and flow rate (Abian et al., 1999). Since its first successful application in the separation of tryptic protein digests, nanoLCMS is nowadays very popular in the field of proteomics (Shen et al., 2002). Metabolomics and lipidomics on the other hand mainly rely on normal flow LCMS. Only a few studies demonstrated the use of nanoflow for metabolomics applications. Polar metabolomes were shown to be retained using self packed capillary columns filled with stationary material targeting either cationic or anionic metabolomes (Myint et al., 2009a; 2009b). Another protocol employs ion pairing chromatography with commercially available nanoLC columns (Kiefer et al., 2010). For unpolar metabolomes or lipidomes, a reversed phase capillary LCMS with self packed columns was described (Gao et al., 2012). All of these protocols share a very high sensitivity with limits of detection values in the low femtomolar range or below. Since each of these methods target only a certain class of metabolites, a combination of different LCMS methods would be necessary in order to cover the metabolome as comprehensive as possible. Combining different methods however increases total instrument time of each sample. All of the described methods employ gradients with a total run time of at least 60 minutes. With the advent of particle sized below 2 μm and ultrahigh pressure chromatography (UPLC), fast chromatographic runs without losses in resolution are possible (Plumb et al., 2004). In this study we present two fast and sensitive, capLCMS methods based on UPLC chromatography to cover a large range of chemically different metabolite classes. Fast gradients employing reversed phase (RP) for lipidome and hydrophilic interaction chromatography (HILIC) for metabolome analysis are used and we demonstrate a comprehensive analysis of both non polar and polar compounds within 35 minutes of total run time.

1.3 Material and Methods

1.3.1 Manufacturing of capillary columns

Polyimide coated fused silica tubing with an internal diameter of 200 μm and an outer diameter of 360 μm were used for capillary LCMS analysis. The silica tubings were obtained from BGB Analytik AG, Switzerland. The capillaries were tapered using a laser capillary puller and chromatography material was filled under nitrogen gas pressure at 30 bar (Myint et al., 2009a).

1.3.2 Selection of HPLC/UPLC materials for the evaluation of polar metabolites

Four different HPLC materials and mobile phases for the analysis of polar metabolites were selected in order to test their applicability in capillary-scale LCMS analysis: 1) Imtakt Scherzo SM-C18 mixed mode column material, 3 μm, 25 mm with 1.5 mM Acetic acid, 5% Methanol as mobile phase A and 50% isopropanol, 100 mM ammonium acetate as mobile phase B. A 10 minute gradient from 0-100% B was employed. 2) Waters BEH amide, 1.7 um, 25 mm with mobile phase A: 50% acetonitrile, 10 mM ammonium acetate and B: 95% acetonitrile, 10 mM ammonium acetate. For both mobile phase buffers, 25% ammonium solution was used to adjust for pH 9. A 20 min gradient from 98%B to 98% A was used. 3) Microsolv Diamond Hydride, 4um, 30 mm was employed using a gradient of 0% to 60% A within 20 minutes, with A=H_2O and B=90% acetonitrile. 10 mM ammonium acetate and 10 mM acetic acid was added to both mobile phases. 4) Waters HSS T3 C18 column, 1.8 um, 15 mm was used for ion pairing chromatography. Solvent A was 1 mM tributylamine, 1.5 mM acetic acid, 5% MeOH, pH 4. Solvent B was methanol. A solvent gradient from 100%A to 0%A within 20 minutes was employed. For the analysis of hydrophobic molecules and lipids, we used a Waters HSS T3 C18 column (1.8 um, 5 mm), using H_2O as mobile phase A and 90% isopropanol/10% acetonitrile as mobile phase B. 5 mM ammonium acetate was added to both mobile phases. A 10 min elution gradient from 100% to 0% A was used.

1.3.3 LCMS analysis

Analyses were performed using a nanoACQUITY system coupled to a Synapt G2HD mass spectrometer (Waters Corp., Milford, USA). For mass spectrometry, all analyses for polar metabolites were done in negative mode using 1.2 kV capillary voltage, 30 V sampling cone voltage and 3 V extraction cone voltage. Source temperature was set to 100 °C and sheet gas was applied. For lipids and hydrophobic metabolites we used positive mode acquisition, using the same settings as in negative mode with the exception of 3.0 kV capillary voltage. MS Data was acquired in MS^E mode. Trap Collision Energy was set to 20-40 V ramp. Mass analysis was set to resolution mode (FWHM ~20000) and a m/z window between 50-1200 was selected. Before injection, samples were diluted 1:5 using the respective initial solvent composition.

1.3.4 Metabolite standards

For the characterization of the different HPLC column materials we evaluated the performance of a standard metabolite mixture, consisting of 89 compounds at a concentration of 100 uM each (Table A-2). For the evaluation of the finally proposed method, we selected 200 small molecule standards, covering different metabolite classes from very polar to non-polar compounds. For LCMS analysis of all standards, we designed 20 mixtures containing 10 metabolites each. A dilution set of the 20 mixtures was created with a concentration range between 100 uM and 100 fM. A list of all standards and their arrangement within the 20 mixes can be seen in (Table 1.3). All standard substances were obtained from Sigma Aldrich in the highest availiable purity.

1.3.5 Metabolite extracts from colorectal tissue

In order to investigate technical and biological variations in real samples, we collected biopsies from colorectal tissue. Five specimens of normal mucosa and cancer were collected from two patients, respectively. Another five specimens of normal mucosa were collected from two additional patients. To evaluated the reproducibility of the method, we pooled aliquots of all samples together. This pooled sample was analysed ten times. Additionally, each individual sample was analysed in three technical replicates. Ultrapure methanol was added to a mixture of internal standards (Appendix Table A-1).

For the extraction of metabolites, 100 ul of internal standard containing methanol was added to 5 mg of each biopsy and the tissue was homogenized by means of a glass homogenizer. The supernatant was collected and subsequently dried in a speedvac vacuum concentrator. The dry metabolite pellet was resuspended in ultrapure water, diluted 1:5 with injection solvent (50 mM NH4 Acetate in 90% ACN, 9% MeOH and 1% H2O, pH9) and transferred to autosampler vials for LCMS analysis.

1.3.6 Data analysis

Each Waters raw file was centroided and converted to vendor independent netCDF data format using MassLynx 4.1 and DataBridge Software from Waters. The netCDF data were processed using the open-source data processing software xcms (Smith et al., 2006). The analytical standards were processed using the matched filter algorithm, with the parameter settings snthresh=3 and step=0.01. From the genereted peak list we retrieved the parameters "into" (Peak Area), "sn" (signal-to-noise ration), "mz" (estimated accurate mass) and "rt" (estimated retention time average). For the processing of the colorectal tissue dataset we used xcms with the centWave algorithm (Tautenhahn et al., 2008) and the following parameter settings to generate peak tables: centWave: ppm=30 and snthresh=3, retcor=„obiwarp", profStep=0.5 and group=„nearest", mzCheck=0.01, rtCheck=15. For the calculation of relative standard deviations (RSD), no further data transformation was performed. For a between group analysis (bga) of the data, the peak tables were log(2)-transformed in advance. The bga function was called from the made4 package using the option „pca" (Culhane et al., 2005).

1.4 Results and Discussion

1.4.1 Selection of chromatographic material for very polar metabolites

Our aim was to develop a metabolomics method which is sensitive enough to cover a wide range of substance classes from a small amount of sample. To achieve this goal of high sensitivity, our strategy was to employ capillary-scale liquid-chromatography mass spectrometry (capLCMS) instead of normal-flow LCMS. It was shown that the decrease of inner column diameter from 1 mm to 250 um can lead to a significant increase of mass signal

intensity (Abian et al., 1999; Richard T Gallagher et al., 2003). Because of the chemically heterogeneous nature of the metabolome we decided to focus on the development of two different analytical methods: One method for very non-polar metabolites and lipids, and one method for very polar compounds. Reversed phase chromatography is commonly found in metabolomics applications and is popular due to very good separation performance but is restricted to hydrophobic metabolites or lipids (Yanes et al., 2011). To retain polar metabolomes by LCMS, different strategies were described in the literature. In order to find the most promising analytical technique for our capillary LCMS system, we evaluated four different chromatographic materials which were described for the analysis of polar metabolomes using normal flow LCMS: 1) HILIC chromatography employing a basic elution buffer system (modified from Bajad et al., 2006). Instead of the originally described material (Luna NH2 from Phenomenex, 5um particle size) we chose a similiar material with 1.7um particles, Waters BEH Amide. 2) A multimodal column material, Scherzo SM-C18 (3 um). This material combines cation and anion exchange with reverse-phase chromatography (modified from Yanes et al., 2011). 3) Reversed phase ion pairing chromatography (modified from Buescher et al., 2010). For this chromatography type, ion pairing reagents are added as a linker between polar analytes and non-polar stationary phase. For this method, we used Waters HSS T3 material (1.8 um) with tributylamine (TBA). 4) Aqueous normal phase chromatography (modified from Pesek et al., 2008). This type of chromatography employs hydride modified silica particles. A diamond hydride column (4 um) from Cogent was used. Each column material was filled into a home made tapered fused silica capillary tip.

To evaluate the analytical performance between these four different LCMS systems, we analyzed a mixture of 89 polar metabolite standards (See Appendix Table A-2) at a concentration of 100 uM each (Figure 1.1). For each detected metabolite, the following properties were calculated: Peak area, retention factor k and signal-to-noise ratio (S/N). The retention factor k characterizes the migration of analyte in relation to the solvent front. k was calculated based on the formula $k = \frac{t_R - t_0}{t_0}$, where t_R is the retention time of the analyte and t_0 is the elution time of the solvent front (Snyder et al., 2011). We estimated the solvent front based on the peak list produced by XCMS, where

t_0 was defined as the retention time of the earliest cohort of detected peaks. The most metabolites (70 of 89) were detected using Waters BEH Amide material, followed by Waters HSS T3+TBA (67), Imtakt Scherzo SM-C18 (64) and Cogent Diamond Hydride (58). In terms of signal sensitivity, BEH Amide showed a similiar overall signal intensity compared to the HSS T3+TBA and together they outperform Diamond Hydride and Scherzo SM-C18 (Figure 1.1a). However, the signal-to-noise ratio of BEH Amide chromatogtaphy is better for most compounds compared to the other materials, indicating a better overall sensitivity of the BEH Amide chromatography (Figure 1.1 c). The best overall chromatographic behaviour was seen for BEH Amide or the Diamond Hydride material, indicated by a factor *k* higher than 1 for the most analytes (Figure 1.1 b). About a half of the compounds analyzed with Scherzo SM-C18 or HSS T3+TBA chromatography have *k* values around 0, indicating that these metabolites elute in the void volume.

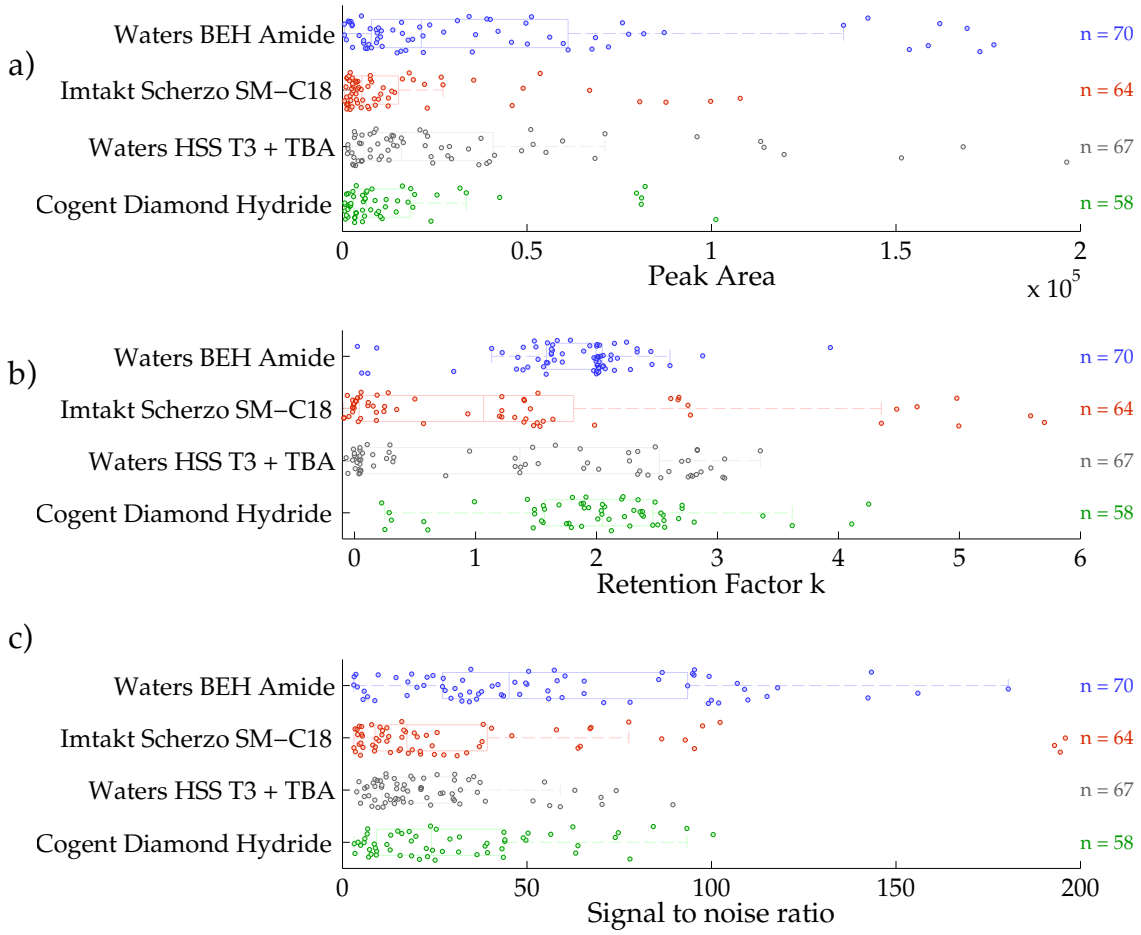


Figure 1.1: Evaluation of chromatography material. Peak Area (a), Retention Factor *k* (b) and Signal to noise ratio (c) was calculated for 89 selected metabolite standards (Appendix Table A-2). The standard compounds were analyzed as mixture at a concentration of 100 uM each. Number n=x indicates the number of detected metabolites for each method.

1.4.2 Method optimization for HILIC chromatography

Based on the above described material evaluation we decided to further work with HILIC chromatography (BEH Amide) for the analysis of polar metabolites and attempted to optimize the chromatographic conditions for this material. The most critical parameter during this process was to find an optimal concentration range of ammonium acetate. This salt is essential for analyte retention on the HILIC column but high concentrations are unwanted because of ion suppression effects. Many standard metabolites were clearly visible in the extracted ion chromatograms (EIC) but not in the base peak intensity chromatogram (BPI) when 10 mM ammonium acetate was added to the elution buffer (Figure 1.2, right panel). This concentration range was recommended for normal flow HILIC LCMS by the manufacturer but the absence of metabolite signals in the BPI indicates ion suppression effects.

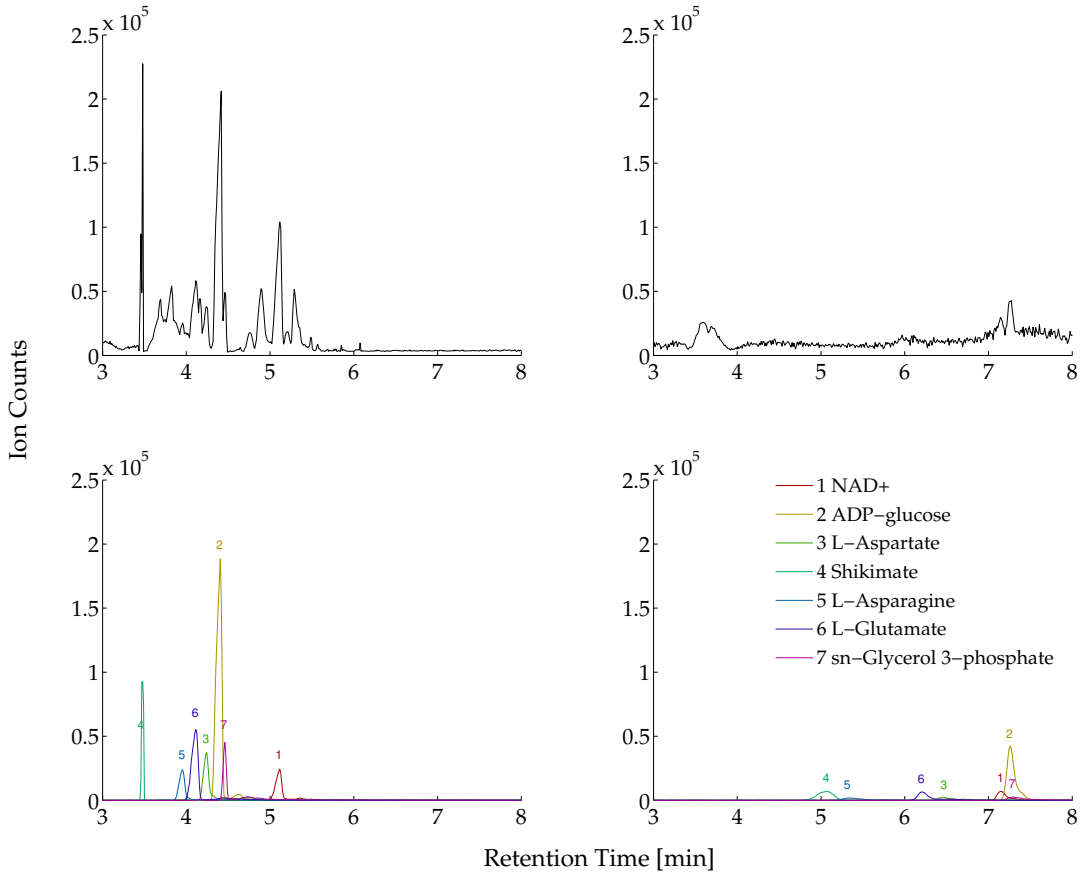


Figure 1.2: Ion suppression and retention time of HILIC chromatography is affected by the ammonium acetate (NH4Ac) concentration in the elution solvents. Upper panel: Base peak intensity chromatograms (BPI) of 29 standard compounds. Each compound was injected at a concentration of 10 uM. Lower Panel: Extracted ion chromatograms of 7 selected standard metabolites. Left Panel: HILIC chromatography with 0.5 mM NH4Ac in the elution buffer, Right Panel: Chromatography with 10 mM NH4Ac.

When lowering the concentration of ammonium acetate in the elution buffer to 0.5 mM, the standard metabolite signals became clearly visible in the BPI (Figure 1.2, left panel). In addition, signal intensities for most analytes increased more than five-fold. The decrease in ammonium acetate concentration expectedly decreased the retention time of each standard compound. When the mixture of 29 standard metabolites contained low concentrations of ammonium acetate in the injection vial, nearly all metabolites eluted together with the solvent front (Figure 1.3, top panel). Only with the increase of ammonium acetate to a high concentration in the injection vial we could achieve a good chromatographic separation while keeping the ammonium acetate concentration low in the elution buffer (Figure 1.3, bottom panel).

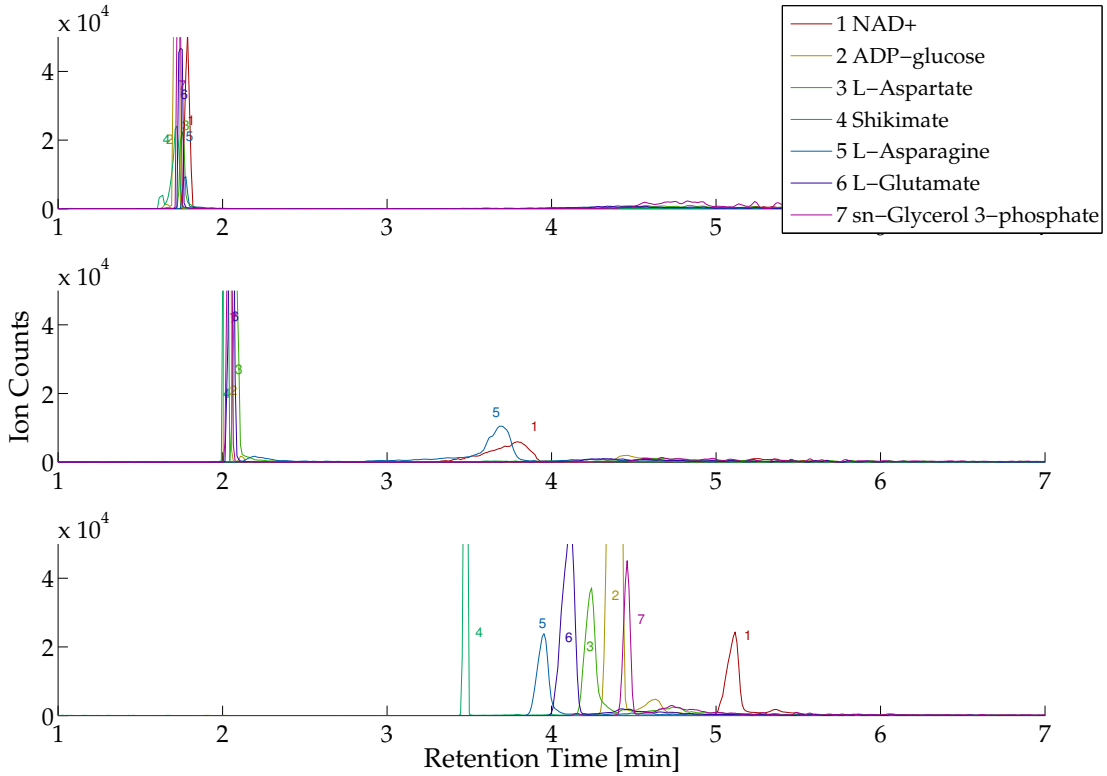


Figure 1.3 A high ammonium acetate (NH4Ac) concentration is essential for the binding and separation of analytes on the HILIC column (BEH amide). Extracted ion chromatograms (EIC) for 7 selected metabolites are shown. Injection solvent with different NH4Ac concentrations were used to dilute the metabolite mixture. Top EIC: 0.5 mM NH4Ac, middle EIC: 5 mM NH4Ac, bottom EIC: 50 mM NH4Ac. The elution buffers contained 0.5 mM NH4Ac.

Finally, we achieve a good compromise of low ion suppression and analyte retention by reducing the ammonium acetate concentration in the elution buffer while maintaining a high amount of ammonium acetate in the injection

solvent. For the analysis of nonpolar metabolites and lipids, we modified a reversed phase LCMS method based on (Castro-Perez et al., 2010). This method was originally developed for lipids and contains high concentrations of organic solvents in both elution buffers. We removed the organic solvent portion during the initial elution gradient in order to increase the chromatographic range for slightly hydrophobic substances.

1.4.3 Characterization of standard compounds

The final capillary LCMS methods for the analysis of polar and unpolar metabolomes are summarized in (Table 1.1). Both methods involve small particles with diametes below 2 um and are therefore suitable for ultrahigh pressure liquid chromatography (UHPLC). Each gradient is completed after only 14 (HILIC) or 20 (reversed phase) minutes and the methods can be used for high throughput analysis of about 100 (HILIC) or 72 (reversed phase) samples per day. We characterized the analytical performance of the two methods with 190 metabolite standards (Table 1.3). These standards were selected to cover a broad range of human metabolic pathways in the KEGG database (Ogata et al., 1999). Each standard metabolite was analyzed in dilution steps from 20 pmol to 0.2 fmol. Limit of detection (LOD) value for each substance was obtained by choosing the dilution step where no signal was observed or where the signal-to-noise ratio (SNR) was below 3.

Table 1.1: Capillary LCMS method overview

Method Conditions	**Polar Metabolome**			**Non-polar metabolome/lipidome**		
Column material	BEH Amide (Waters)			HSS T3 (Waters)		
Particle Diameter	1.8 um			1.7 um		
Column length	15 cm			5 cm		
Column diameter	200 um			200 um		
Injection solvent (1:5 diluted w/ analyte)	90% acetonitrile, 9% methanol, 1% water, 50 mM ammonium acetate			Solvent A		
Solvent A:	0.5 mM ammonium acetate, pH 9			5 mM ammonium formate		
Solvent B:	95% acetonitrile, 0.5 mM ammonium acetate, pH 9			90% isopropanol, 10% acetonitrile, 5 mM ammonium formate		
MS polarity	negative			positive		
Gradient conditions:	Minutes	Composition B [%]	Flow rate [ul/min]	Minutes	Composition B [%]	Flow rate [ul/min]
	0.00	90.0	6.0	0.00	0	4
	10.00	50.0	4.0	10.00	100	3
	12.00	50.0	4.0	15.00	100	3
	12.01	90.0	4.0	15.01	0	3
	14.00	90.0	4.0	20.00	0	3

From the 190 metabolite standards, 167 could be detected at least at 20 pmol concentration. From these, 123 were well retained ($k > 1$) on the HILIC column and 56 on the reversed phase column. The high amount of metabolites with good separation efficiency on the HILIC column highlights the importance of a non-polar stationary phase for the analysis of central carbon metabolism. However, for most of the rather non-polar metabolites the reversed phase column is superior to the HILIC column (Figure 1.4). Very polar metabolites such as ADP-ribose retain on the HILIC column but elute within the void volume of the reversed phase column. Complementary, non-polar compounds suchas taurocholic acid are only retained on the reversed phase column. For most of the metabolite standards, the theoretically calculated logP value determined the fate of the respective analyte on either of both columns (Table 1.2). Polar metabolites with logP below zero retained preferantially well on the HILIC column whereas non-polar compounds around and above zero retained better on the reveresed phase column.

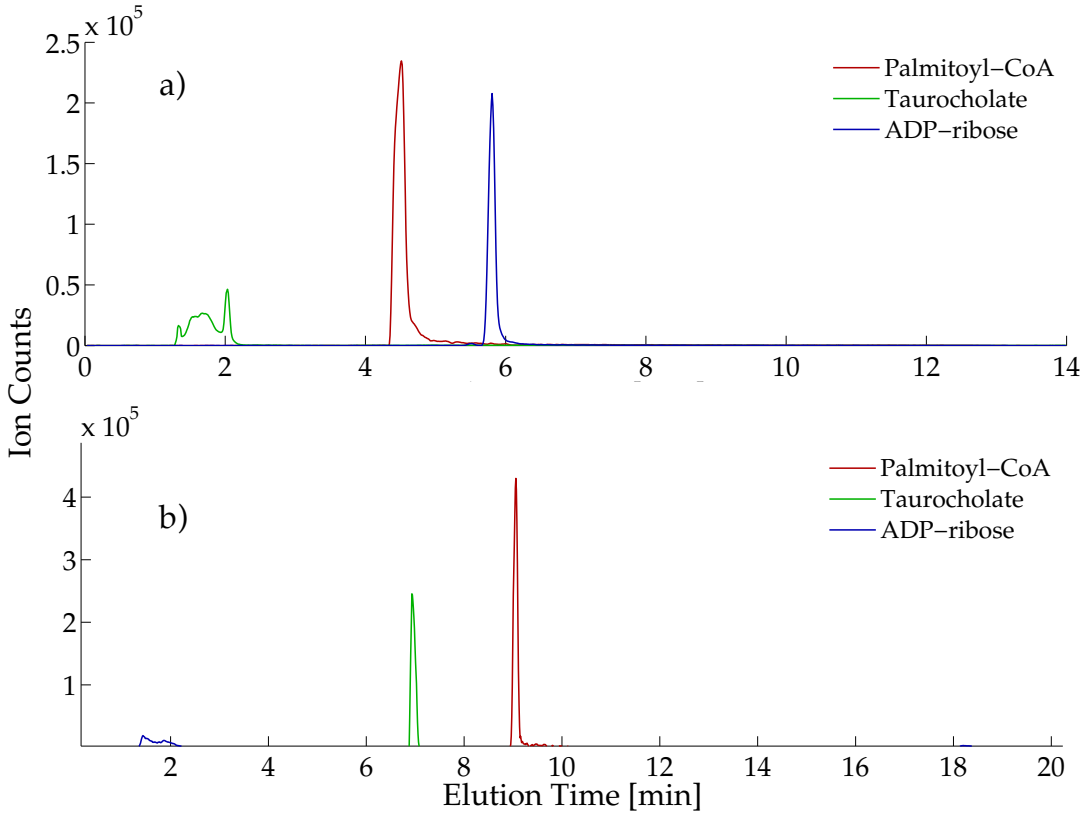


Figure 1.4: Comparison between a) HILIC and b) reversed phase chromatography based separation of Palmitoyl-CoA, Taurocholate and ADP-Ribose. Very polar ADP-ribose is well retained using HILIC material BEH amide but elutes within the void volumn of a reversed phase column. Complementary, taurocholic acid retains well on the reversed phase column but not on the HILIC column. Amphipathic Palmitoyl-CoA can be seperated on both types of chromatography.

From all 167 detected metabolites, 152 could be retained on either HILIC or reversed phase chromatography and both methods nicely complement each other. For 63% of the metabolite standards (102) we could measure limit of detection values (LOD) in the femtomolar range (Table 1.3). The achieved sensitivity in full scan mode is therefore comparable to contemporary targeted LCMS methods where triple quadrupole mass spectrometer and normal flow chromatography are used (Luo et al., 2007; Buescher et al., 2010). This improved sensitivity might allow for improved detection of unknown peaks in a metabolic profiling experiment while maintaining the sensitivity for a targeted metabolome quantification. Despite a relatively short running time and a fast gradient, the capillary LCMS methods provide a good chromatographic resolution. A challenge in metabolomics is the separation of biologically relevant isomers and mass spectrometry alone does not allow for unambiguous annotation of these compounds. Especially for central metabolism it is of importance to distiguish between different forms of sugar phosphates (Buescher et al., 2010). For most of the sugar phosphates such as glucose-6-phosphate, fructose-6-phosphate and mannose-6-phosphate, baseline separation can be nearly achieved by using HILIC chromatography (Figure 1.5).

Table 1.2: Retention behaviour of compounds with different polarity on a HILIC or reversed phase capillary LC column. Retention factor *k* (migration distance of analyte compared to solvent front) is given for hydrophobic (LogP > 0) or polar (LogP < 0) metabolites.

Name	logP	*k* HILIC	*k* Reversed Phase
2-Phenylacetamide	0.64	0.1	1.6
3-Methoxytyramine	-0.04	0.2	1.1
4-Pyridoxate	-0.08	0.0	1.2
Homogentisate	0.81	0.0	4.1
Indole-3-acetate	1.87	0.1	1.8
Melatonin	1.42	0.0	2.2
N-Acetylserotonin	0.98	0.0	1.5
N-Methyltryptamine	2.02	0.2	1.5
Phenylacetaldehyde	1.75	0.0	4.0
Taurocholate	0.79	0.0	3.6
ADP	-1.65	3.9	0.0
ADP-ribose	-1.8	2.7	0.0
Asparagine	-3.36	2.3	0.1
Citrate	-1.33	3.8	0.0
Ethanolamine phosphate	-2.5	3.5	0.0
Fructose 6-phosphate	-2.11	3.2	-0.1
GDP	-1.51	4.3	-0.1
Glucarate	-3.6	2.8	-0.1
Glutamate	-3.54	2.3	0.1
Glyceraldehyde 3-phosphate	-1.69	3.2	0.0

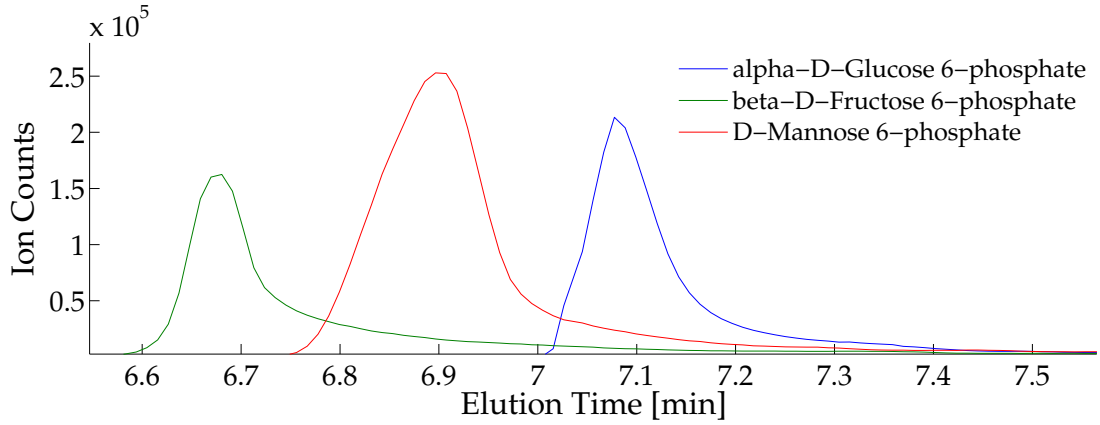


Figure 1.5: Chromatographic performance of very polar metabolites on a capillary UPLC column filled with BEH amide particles. 4 pmol of each substance was injected on column. Biologically relevant hexose-phosphate isomers can be seperated from each other by chromatography.

1.4.4 Application of capillary LCMS methods to metabolite extracts of low abundant biological material

The focus of this study was to find a sensitive method for a comprehensive metabolomics analysis of low abundant biological material, for example from small biopsies (<5mg). In order to test the proposed methods with a real application, we demonstrate a metabolome and lipidome analysis from colorectal tissue. A low amount of fresh biopsies (3-5 mg) from surgically removed colon tissue was collected. Normal mucosa and cancer tissue were chosen to investigate metabolome differences between both disease states. Metabolome and lipidome extracts were prepared by using a single methanol extraction step (Römisch-Margl et al., 2011). Accurate masses were matched to the KEGG database for metabolites (Ogata et al., 1999) or to the HMDB database for lipids (Wishart et al., 2013). The HMDB database was filtered for the super class „lipids“ prior to the search. 378 masses of the HILIC and 107 masses of the reversed phase method could be annotated to 845 unique metabolite masses in the KEGG database. Additionally, 102 masses in the HILIC dataset and 688 in the reversed phase dataset were annotated to the lipid subset of the HMDB database, (Figure 1.6). Together, these results again demonstrate the complementarity of both methods for the analysis of polar metabolomes using HILIC and nonpolar metabolomes and lipidomes using reversed phase capillary chromatography based LCMS.

For the analysis of metabolomics data, especially for a non-targeted experiment, it is often desirable to interpret the data on the basis of altered metabolic pathways. To further characterize the annotated metabolome space in both of our methods, we investigated the coverage of metabolites in human metabolic pathways of the KEGG database (Ogata et al., 1999) (Figure 1.7). Except for the glycan biosynthesis pathway, metabolites of all pathways could be annotated. Notably, most of the metabolites involved in carbohydrate, energy, amino acid and nucleotide metabolism were annotated only in the HILIC dataset whereas metabolites found in the reversed phase dataset were more present in lipid metabolism. For each of these tissue groups (normal mucosa and cancer) we collected 5 replicates to investigate the biological variance. In order to test the technical robustness of the method, we created a pool sample from all extracts and injected this sample after every 6th LCMS run. Additionally, each sample was analyzed in three technical replicates. The median relative standard deviation (RSD) value for the annotated metabolites in the pool sample was 15.1% for the HILIC and 22.4% for the reversed phase method (Figure 1.8). Both methods provide therefore good overall technical reproducibility.

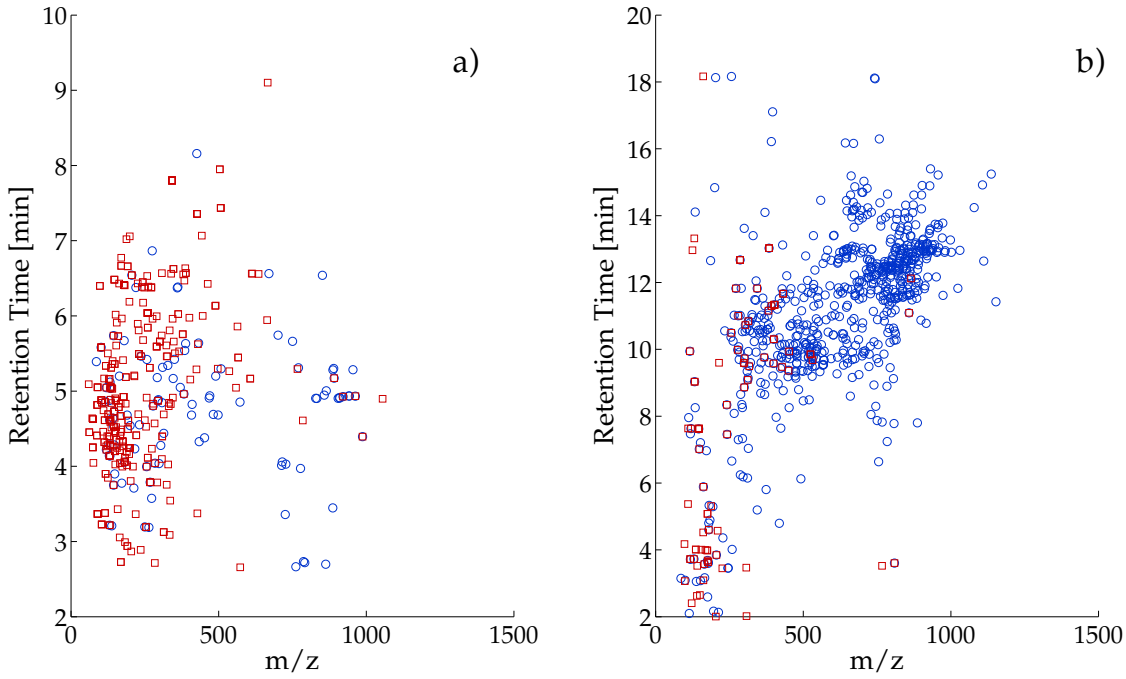


Figure 1.6: Retention behaviour of a colorectal tissue metabolome (red, squares) and lipidome (blue, circles). 100 metabolite/lipid extracts from colorectal normal mucosa or cancer tissue were analyzed on the HILIC (a) or reversed phase (b) capillary LCMS system. Exact masses of a XCMS processed peak table were annotated as metabolites using the KEGG database (Ogata et al., 1999) or as lipids using the human metabolome database (Wishart et al., 2013).

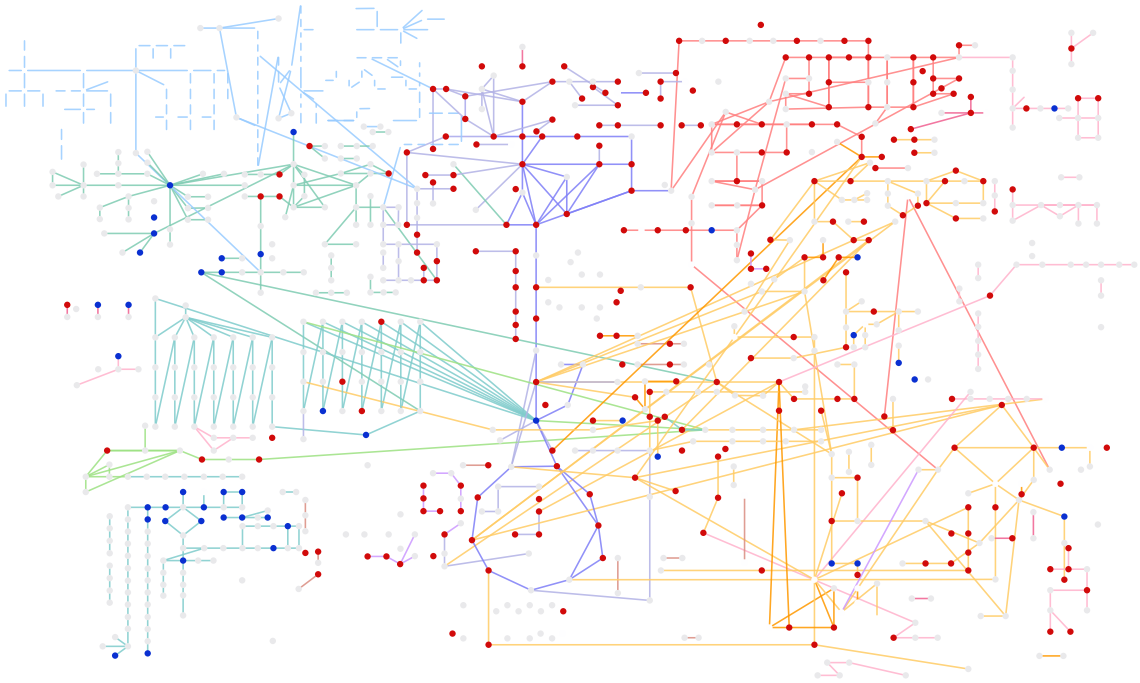


Figure 1.7: Coverage of annotated metabolites from the human metabolic pathways of the KEGG database. Colored lines indicate enzymatic reactions and each color corresponds to a specific metabolic pathway cluster. Green: Lipid metabolism, Dark blue (middle): Carbohydrate metabolism, Purple: Energy metabolism, Yellow: Amino acid metabolism, Red: Nucleotide metabolism, Bright blue (upper left): Glycan biosynthesis. Red dots: Metabolites annotated in the HILIC dataset. Blue dots: Metabolites annotated in the reversed phase dataset. Gray dots: Metabolites not annotated.

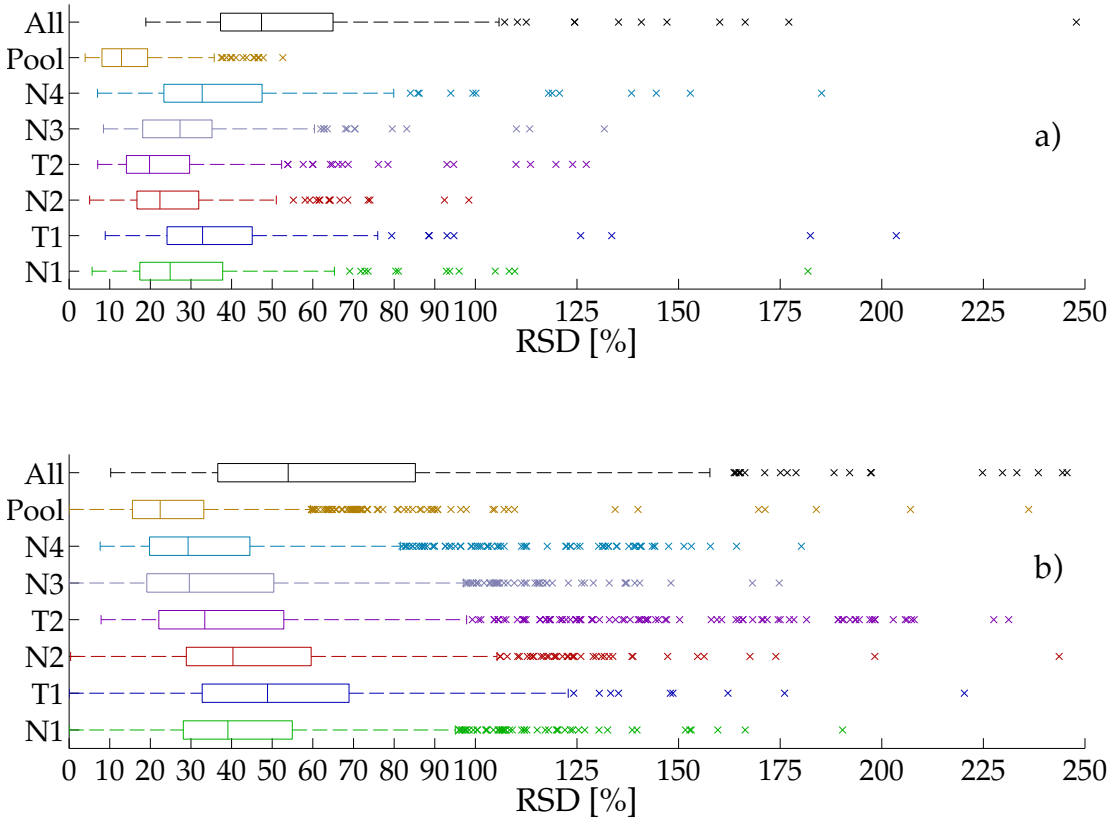


Figure 1.8: Relative standard deviations (RSD in %) of 330 annotated metabolites in the HILIC method (a) and 1100 annotated lipids of the reversed phase method (b). All: RSD values of all samples, Pool = Pooled sample from all extracts. N = normal mucosa, T = tumor. Numbers indicate the patient number.

To further characterize the biological and technical reproducibility of both methods, we performed a between group analysis (BGA) on the annotated metabolome and lipidome data set (Culhane et al., 2002). BGA deals with a common problem in most omics datasets, where more number of variables than cases are available. Instead of decomposing the raw data matrix on each variable (metabolites), a predefined number of groups is used. The BGA was carried out based on principal components analysis (PCA) and the following seven groups were defined: Normal mucosa from four patients (N1, N2, N3, N4), tumor tissue from two patients (T1, T2), and the pool sample (Pool). The first BGA axis clearly discriminates between a normal mucosa and cancer cluster in both datasets, indicating that the highest variation in the metabolome and lipidome is explained by the difference between normal mucosa and cancer (Figure 1.9). The second axis discriminates between the two different tumor tissues whereas the four normal mucosa tissues cluster closely together, suggesting that the metabolome of normal mucosa between different patients is more homogenous than the metabolome between the two tumors. The pool sample cluster is located in the center of both axes and shows only little variation. Also, the technical replicates are very close together with a similiar small variation compared to the pool samples. Again this demonstrates a good technical robustness of both methods. Within the clusters, especially in the tumor 1 (T1), there is a clear intra-group variation observable which exceeds the technical variation. This phenomenon might reflect the fact that biopsies of colorectal tissue consist of a very heterogenous mixture of different cell types. Consequently, the metabolome of biopsies from the same region but not exactly the same location may vary. However, the clear separation of the cancer and normal mucosa clusters on the first axis indicates that the metabolome changes between normal cells and cancer cells are reflected well in the extracts of the tissue biopsies.

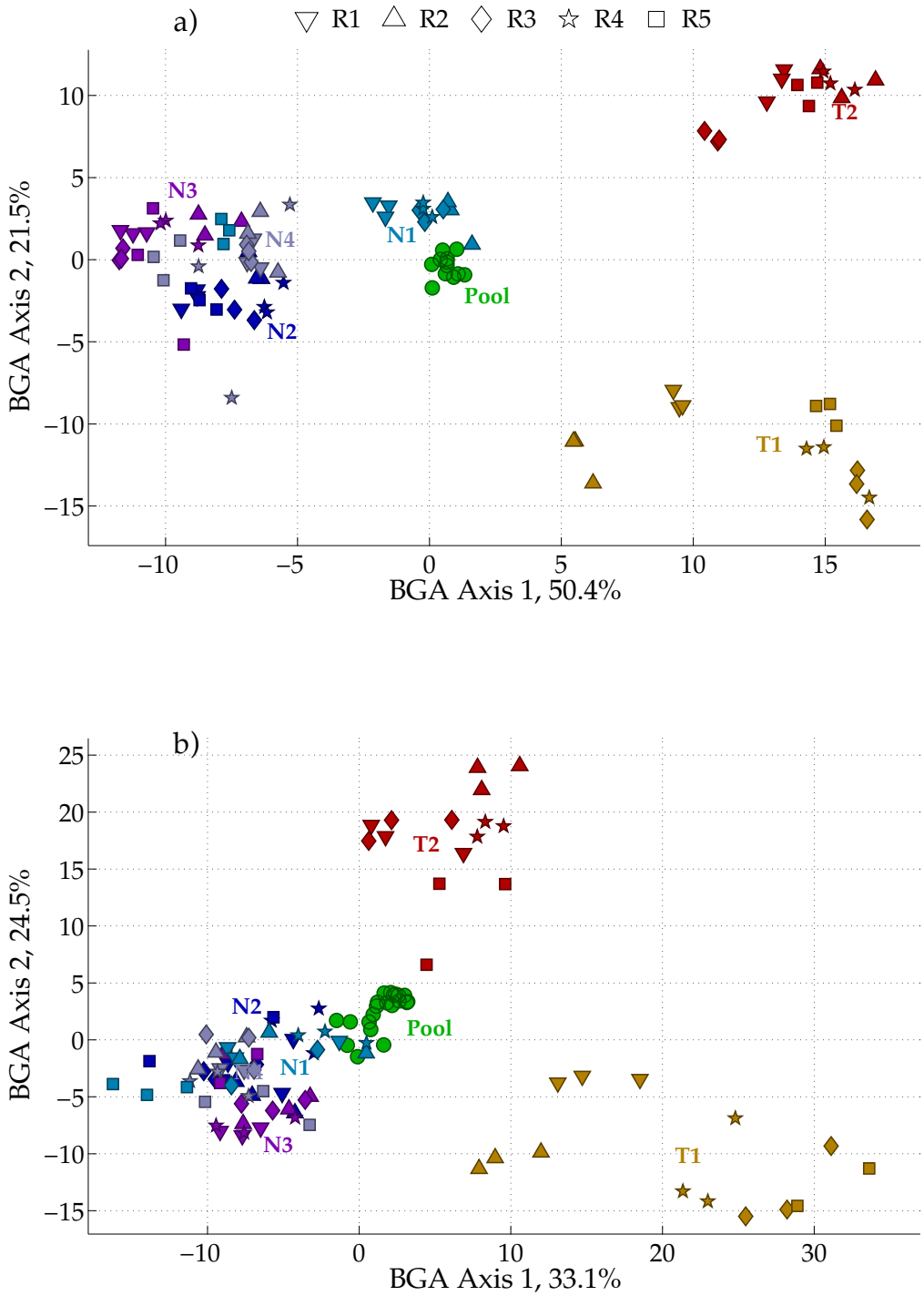


Figure 1.9: Between group analysis (BGA) of colorectal tissue metabolome using capillary HILIC LCMS (a) and lipidome using capillary RP LCMS (b). Score plots of the BGA are shown. Colors indicate each biopsy cohort (normal mucosa or tumor). Pool = Pooled sample from all extracts. N=normal mucosa, T=tumor. Numbers indicate the number of patient or biological replicate. Technical replicates are indicated by a unique symbol.

1.5 Conclusion

We present two methods based on capillary LCMS for a sensitive quantification of metabolome and lipidome extracts from a small amount of

biological sample. Two complementary methods were chosen to address the very heterogenous polarity nature of the metabolome. After testing different column chemistries, a HILIC method was chosen for very polar metabolites. For lipids and non polar metabolites a reversed phase method was employed. For 167 standard compounds we report limit of detection (LOD) values in the low picomolar to femtomolar range. As a result, our method is suitable for non-targeted metabolomics by covering a broad range of metabolite masses in full-scan mode while providing high sensitivity using capillary flow rates. The achieved sensitivity is comparable to contemporary targeted LCMS methods using triple quadrupole mass spectrometers and normal flow chromatography (Luo et al., 2007; Buescher et al., 2010). Because of UPLC particles with sizes below 2 μm and a fast chromatographic gradient, one analytical run takes only 14 minutes for the HILIC and 20 minutes for the reversed phase method. Despite such a fast gradient, a good chromatographic separation can still be achieved. We demonstrate the applicability of the presented method in a metabolomics scenario with metabolome and lipidome extracts of colorectal tissue biopsies where only minimal amounts of biological sample are available (< 5 mg tissue). About a half of the metabolites in the KEGG database and additionally about 700 lipids could be annotated based on their accurate mass. For most of the quantified metabolites we achieve good technical reproducibility as indicated by RSD values between 15 and 20% RSD. We recommend the described set of capillary LCMS methods primarily for high-throughput metabolomics experiments relying on a few hundreds of low volume samples. As the methods feature low solvent consumption and perfect compatibility to nanoESI mass spectrometry platforms, as in use in proteomics, we recommend consideration of this methods also in cases where sample size does not matter but platform flexibility is an asset.

Table 1.3: Standard compound characteristics and performance on the negative mode hydrophilic interaction chromatography (HILIC (-)) and the positive mode reversed phase (RP (+)) method. LOD = limit of detection. t_R = retention time, k = retention factor $k = \frac{t_R - t_0}{t_0}$ with t_0 = retention time of the solvent front.

Metabolite Information					HILIC (-)			RP (+)		
Name	**KEGG ID**	**Formula**	**Mass**	**logP**	**LOD**	t_R	k	**LOD**	t_R	k
1-Methylnicotinamide	C02918	C7H9N2O	137.072	-3.72				200 fmol	1.7	0.2
1.3-Diaminopropane	C00986	C3H10N2	74.0844	-1				20 pmol	5.1	2.4
2-Aminoethylphosphonate	C03557	C2H8NO3P	125.024	-1.55	20 pmol	7.4	3.7	20 pmol	1.7	0.1
2-Deoxy-D-ribose 1-phosphate	C00672	C5H11O7P	214.024	-2.14	200 fmol	6.2	2.9			
2-Oxobutanoate	C00109	C4H6O3	102.032	0.07	20 pmol	1.8	0.1	20 pmol	7.6	4.1
2-Oxoglutarate	C00026	C5H6O5	146.022	-0.6	2 pmol	4.6	1.9			
2-Phenylacetamide	C02505	C8H9NO	135.068	0.64	2 pmol	1.7	0.1	20 fmol	3.9	1.6
2-Phospho-D-glycerate	C00631	C3H7O7P	185.993	-2.24	20 pmol	7.2	3.5	20 pmol	1.5	0.0
2.3-Bisphospho-D-glycerate	C01159	C3H8O10P2	265.959	-1.55				2 pmol	1.4	-0.1
3-Aminoisobutyric acid	C05145	C4H9NO2	103.063	-2.97	2 pmol	4.8	2.0	2 pmol	1.6	0.1
3-Aminopropiononitrile	C05670	C3H6N2	70.0531	-0.9						
3-Hydroxy-L-kynurenine	C03227	C10H12N2O4	224.08	-2.09	2 pmol	3.9	1.5	2 pmol	1.7	0.1
3-Hydroxybutanoate	C01089	C4H8O3	104.047	-0.5	2 pmol	3.3	1.0			
3-Methoxy-4-hydroxymandelate	C05584	C9H10O5	198.053	0	200 fmol	1.6	0.0			
3-Methoxytyramine	C05587	C9H13NO2	167.095	-0.04	20 pmol	1.9	0.2	2 pmol	3.2	1.1
3-Ureidopropionate	C02642	C4H8N2O3	132.054	-0.98	2 pmol	4.0	1.5			
3.4-Dihydroxy-L-phenylalanine	C00355	C9H11NO4	197.069	-2.32	20 pmol	4.3	1.7	2 pmol	1.7	0.2
3.4-Dihydroxymandelate	C05580	C8H8O5	184.037	0.72	20 pmol	2.7	0.7	2 pmol	1.2	-0.2
4-Aminobutanoate	C00334	C4H9NO2	103.063	-2.99	2 pmol	5.6	2.5	2 pmol	1.5	0.0
4-Pyridoxate	C00847	C8H9NO4	183.053	-0.08	2 fmol	1.6	0.0	200 fmol	3.3	1.2
5-Amino-4-imidazolecarboxyamide	C04051	C4H6N4O	126.054	-1.16	2 pmol	1.7	0.0	20 pmol	8.4	4.6
5-Diphosphomevalonate	C01143	C6H14O10P2	308.006	-1.17	2 pmol	7.6	3.7	20 pmol	1.3	-0.1
5-Formyltetrahydrofolate	C03479	C20H23N7O7	473.166	-0.46	0.2 fmol	5.6	2.5	0.2 fmol	3.4	1.3
5-Hydroxyindoleacetate	C05635	C10H9NO3	191.058	1.28						
5-Methoxyindoleacetate	C05660	C11H11NO3	205.074	1.76						
5-Methyltetrahydrofolate	C00440	C20H25N7O6	459.187	-1.26	200 fmol	6.3	2.9	20 fmol	3.4	1.3
5-Phospho-alpha-D-ribose 1-diphos	C00119	C5H13O14P3	389.952	-6				20 pmol	1.3	-0.1
5-Phosphomevalonate	C01107	C6H13O7P	228.04	-2.18	2 pmol	6.4	3.0	20 pmol	1.7	0.1
6-Deoxy-L-galactose	C01019	C6H12O5	164.069	-2.39	20 pmol	2.8	0.8			
Acetoacetate	C00164	C4H6O3	102.032	-0.47	20 pmol	2.8	0.8	2 pmol	7.6	4.1
Acetyl phosphate	C00227	C2H5O5P	139.988	-0.92	200 fmol	5.7	2.5			
Acetyl-CoA	C00024	C23H38N7O17P3S	809.126	-0.58	2 fmol	6.1	2.8	200 fmol	3.4	1.3
Adenine	C00147	C5H5N5	135.055	-0.38	20 pmol	7.2	3.5	2 pmol	2.1	0.4
Adenosine	C00212	C10H13N5O4	267.097	-1.21	200 fmol	1.8	0.1	20 fmol	3.4	1.3
Adenosine 3.5-bisphosphate	C00054	C10H15N5O10P2	427.029	-1.63	0.2 fmol	7.2	3.5	2 pmol	2.0	0.3
Adenylyl sulfate	C00224	C10H14N5O10PS	427.02	-1.64	2 fmol	4.4	1.8			
ADP	C00008	C10H15N5O10P2	427.029	-1.65	20 pmol	7.9	3.9	2 pmol	1.6	0.0
ADP-ribose	C00301	C15H23N5O14P2	559.072	-1.8	20 fmol	5.9	2.7	200 fmol	1.5	0.0
Adrenaline	C00788	C9H13NO3	183.09	-0.82						
Agmatine	C00179	C5H14N4	130.122	-1.01						
Alanine	C00099	C3H7NO2	89.0477	-3.26	2 pmol	5.3	2.3	20 pmol	1.4	0.0
Alanine	C00041	C3H7NO2	89.0477	-3.05	2 pmol	4.6	1.9	20 pmol	1.5	0.0
Allantoate	C00499	C4H8N4O4	176.055	-2.12	200 fmol	4.5	1.8	0.2 fmol	7.0	3.7
Allantoin	C02350	C4H6N4O3	158.044	-1.95	200 fmol	2.0	0.3			
Arabinose	C00259	C5H10O5	150.053	-1.9	2 pmol	3.1	0.9			
Arabitol	C00532	C5H12O5	152.069	-3.2	2 pmol	3.6	1.2	2 pmol	7.6	4.1
Ascorbate	C00072	C6H8O6	176.032	-0.5	20 pmol	4.2	1.6			
Asparagine	C00152	C4H8N2O3	132.054	-3.36	200 fmol	5.3	2.3	2 pmol	1.6	0.1
Aspartate	C00402	C4H7NO4	133.038	-3.52	2 fmol	5.7	2.5	200 fmol	1.3	-0.1
ATP	C00002	C10H16N5O13P3	506.996	-0.84	200 fmol	7.3	3.6	20 pmol	##	###
Biocytin	C05552	C16H28N4O4S	372.183	-0.69	20 fmol	4.6	1.9	20 fmol	3.5	1.3
Biotin	C00120	C10H16N2O3S	244.088	0.39						
Butanal	C01412	C4H8O	72.0575	1.1				0.2 fmol	3.8	1.5
Butanoyl-CoA	C00136	C25H42N7O17P3S	837.157	-0.22	200 fmol	5.5	2.4	2 pmol	3.7	1.5
Cadaverine	C01672	C5H14N2	102.116	-0.27				20 pmol	2.6	0.7
cis-Aconitate	C00417	C6H6O6	174.016	-0.41	0.2 fmol	6.3	2.9	0.2 fmol	2.1	0.4

Metabolite Information					HILIC (-)			RP (+)		
Name	**KEGG ID**	**Formula**	**Mass**	**logP**	**LOD**	t_R	k	**LOD**	t_R	k
Citrate	C00158	C6H8O7	192.027	-1.33	2 pmol	7.6	3.8	200 fmol	1.4	0.0
CoA	C00010	C21H36N7O16P3S	767.115	-0.61	20 pmol	6.2	2.9	200 fmol	3.3	1.2
Coniferyl alcohol	C00590	C10H12O3	180.079	1.51	2 pmol	1.7	0.1	20 pmol	4.4	1.9
Cysteamine	C01678	C2H7NS	77.0299	0.01						
Cysteine	C00097	C3H7NO2S	121.02	-2.57	2 pmol	6.2	2.9	20 pmol	1.5	0.0
dADP	C00206	C10H15N5O9P2	411.035	-1.57	200 fmol	6.8	3.2	2 pmol	2.1	0.4
dAMP	C00360	C10H14N5O6P	331.068	-2.44	2 fmol	5.9	2.7	2 pmol	2.5	0.7
Decanoyl-CoA	C05274	C31H54N7O17P3S	921.251	0.87	20 fmol	4.7	1.9	2 pmol	7.2	3.8
Deoxyadenosine	C00559	C10H13N5O3	251.102	-0.95	200 fmol	1.7	0.1	20 fmol	3.4	1.3
Deoxyguanosine	C00330	C10H13N5O4	267.097	-1.75	20 fmol	2.0	0.2	200 fmol	3.3	1.2
dGDP	C00361	C10H15N5O10P2	427.029	-1.45	20 pmol	7.8	3.9	2 pmol	1.5	0.0
dGMP	C00362	C10H14N5O7P	347.063	-1.87	0.2 fmol	6.5	3.1	2 pmol	1.8	0.2
dGTP	C00286	C10H16N5O13P3	506.996	-0.61	0.2 fmol	8.1	4.1	2 pmol	1.4	0.0
Dihydrofolate	C00415	C19H21N7O6	443.155	-3.2						
Dihydroorotate	C00337	C5H6N2O4	158.033	-1.7				20 pmol	6.6	3.4
Dimethylallyl diphosphate	C00235	C5H12O7P2	246.006	0.3	0.2 fmol	6.2	2.9			
Dopamine	C03758	C8H11NO2	153.079	-0.4	20 pmol	2.0	0.3	20 pmol	3.3	1.2
Ethanolamine phosphate	C00346	C2H8NO4P	141.019	-2.5	2 pmol	7.2	3.5	2 pmol	1.5	0.0
Ethylene glycol	C01380	C2H6O2	62.0368							
FMN	C00061	C17H21N4O9P	456.105	-0.78	200 fmol	5.6	2.5	200 fmol	3.7	1.5
Fructose	C00095	C6H12O6	180.063	-2.4	20 pmol	4.6	1.9	200 fmol	7.6	4.1
Fructose 1.6-bisphosphate	C05378	C6H14O12P2	339.996	-1.53	2 pmol	7.6	3.8			
Fructose 6-phosphate	C05345	C6H13O9P	260.03	-2.11	0.2 fmol	6.7	3.2	20 pmol	1.4	-0.1
Fucose 1-phosphate	C02985	C6H13O8P	244.035	-1.53	20 fmol	6.5	3.1			
Fumarate	C00122	C4H4O4	116.011	0.21	2 pmol	4.7	1.9	20 pmol	5.0	2.4
Galactose	C00124	C6H12O6	180.063	-2.57	2 pmol	4.2	1.6			
Galactose 1-phosphate	C00446	C6H13O9P	260.03	-2	0.2 fmol	6.9	3.3			
GDP	C00035	C10H15N5O11P2	443.024	-1.51	2 pmol	8.5	4.3	200 fmol	1.4	-0.1
GDP-L-fucose	C00325	C16H25N5O15P2	589.082	-1.69	2 fmol	6.6	3.1	200 fmol	1.5	0.0
GDP-mannose	C00096	C16H25N5O16P2	605.077	-1.76	2 fmol	7.0	3.4	2 pmol	1.4	0.0
Gentisate aldehyde	C05585	C7H6O3	138.032	1.02	20 fmol	1.6	0.0	20 pmol	4.3	1.8
Glucarate	C00818	C6H10O8	210.038	-3.6	200 fmol	6.1	2.8	2 pmol	1.3	-0.1
Gluconic acid	C00257	C6H12O7	196.058	-1.87	200 fmol	5.0	2.1	20 pmol	7.6	4.1
Glucosamine 6-phosphate	C00352	C6H14NO8P	259.046	-2.6	20 pmol	7.4	3.6	2 pmol	1.4	0.0
Glucose	C00031	C6H12O6	180.063	-2.57	2 pmol	4.2	1.6	20 pmol	1.5	0.0
Glucose 6-phosphate	C00668	C6H13O9P	260.03	-2.06	0.2 fmol	7.1	3.4	2 pmol	4.2	1.8
Glucuronate	C00191	C6H10O7	194.043	-2	20 pmol	5.6	2.5			
Glutamate	C00025	C5H9NO4	147.053	-3.54	20 fmol	5.3	2.3	2 pmol	1.6	0.1
Glutamine	C00064	C5H10N2O3	146.069	-3.32	200 fmol	5.3	2.3	20 pmol	1.5	0.0
Glutaryl-CoA	C00527	C26H42N7O19P3S	881.147	-0.48	20 fmol	6.8	3.3	20 fmol	3.4	1.3
Glyceraldehyde 3-phosphate	C00118	C3H7O6P	169.998	-1.69	2 pmol	6.6	3.2	20 pmol	1.5	0.0
Glycerate	C00258	C3H6O4	106.027	-1.72	20 fmol	3.8	1.4			
Glycerol	C00116	C3H8O3	92.0473	-1.93						
Glycerol 3-phosphate	C00093	C3H9O6P	172.014	-1.84	200 fmol	6.1	2.8	2 pmol	1.4	-0.1
Glycine	C00037	C2H5NO2	75.032	-3.34	20 pmol	5.2	2.3	20 pmol	1.5	0.0
Glyoxylate	C00048	C2H2O3	74.0004	-0.59	2 pmol	3.9	1.4			
GTP	C00044	C10H16N5O14P3	522.991	-0.63	200 fmol	8.2	4.1	2 pmol	1.3	-0.1
Guanine	C00242	C5H5N5O	151.049	-0.9	20 pmol	5.1	2.2	20 pmol	1.3	-0.2
Guanosine	C00387	C10H13N5O5	283.092	-2.06	200 fmol	3.0	0.9	20 pmol	2.8	0.8
Hexanoyl-CoA	C05270	C27H46N7O17P3S	865.188	0.07	2 fmol	5.2	2.2	200 fmol	4.6	2.1
Homogentisate	C00544	C8H8O4	168.042	0.81	2 pmol	1.7	0.0	0.2 fmol	7.6	4.1
Homovanillate	C05582	C9H10O4	182.058	1.02	2 pmol	1.6	0.0			
Hypoxanthine	C00262	C5H4N4O	136.039	-0.55	200 fmol	1.8	0.1	200 fmol	2.1	0.4
IDP	C00104	C10H14N4O11P2	428.013	-2.47	20 pmol	7.6	3.7	2 pmol	1.4	0.0
IMP	C00130	C10H16N5O12P3	491.001	-0.66	20 pmol	7.4	3.7			
Indole-3-acetate	C00954	C10H9NO2	175.063	1.87	200 fmol	1.8	0.1	200 fmol	4.3	1.8
Inosine	C00294	C10H12N4O5	268.081	-1.87	20 fmol	1.9	0.2	20 pmol	2.7	0.8
Inositol	C00137	C6H12O6	180.063	-2.08	200 fmol	6.3	3.0			
Isocitrate	C00311	C6H8O7	192.027	-0.35	2 pmol	7.0	3.4	2 pmol	1.4	-0.1
Isopentenyl diphosphate	C00129	C5H12O7P2	246.006	0.04	200 fmol	6.2	2.9	2 pmol	2.0	0.3

Metabolite Information					HILIC (-)			RP (+)		
Name	**KEGG ID**	**Formula**	**Mass**	**logP**	**LOD**	t_R	k	**LOD**	t_R	k
ITP	C00081	C10H15N4O14P3	507.98	-0.67	20 pmol	8.3	4.2	200 fmol	1.3	-0.1
Lactose	C00243	C12H22O11	342.116	-5.03	20 fmol	6.4	3.0			
Lauroyl-CoA	C01832	C33H58N7O17P3S	949.282	1.35	20 fmol	4.5	1.8	2 pmol	7.9	4.2
Malate	C00149	C4H6O5	134.022	-0.87	0.2 fmol	5.5	2.5			
Malonate	C00383	C3H4O4	104.011	-0.6	20 fmol	5.4	2.4	20 pmol	##	###
Malonyl-CoA	C00083	C24H38N7O19P3S	853.116	-0.62	200 fmol	6.9	3.3	2 pmol	3.3	1.2
Maltose	C00208	C12H22O11	342.116	-5.03	0.2 fmol	6.1	2.8			
Mannitol	C00392	C6H14O6	182.079	-2.68	20 fmol	4.2	1.6	20 pmol	1.5	0.0
Mannose	C00159	C6H12O6	180.063	-2.57	200 fmol	6.4	3.0	200 fmol	7.6	4.1
Mannose 1-phosphate	C00636	C6H13O9P	260.03	-2	0.2 fmol	6.9	3.3	2 pmol	1.5	0.0
Mannose 6-phosphate	C00275	C6H13O9P	260.03	-2.06	200 fmol	6.9	3.3	20 pmol	1.7	0.1
Melatonin	C01598	C13H16N2O2	232.121	1.42	200 fmol	1.6	0.0	20 fmol	4.9	2.2
Methylglyoxal	C00546	C3H4O2	72.0211	-0.38	200 fmol	2.8	0.8			
Methylmalonate	C02170	C4H6O4	118.027	0.17	20 pmol	4.8	2.0	20 pmol	3.4	1.3
N-(L-Arginino)succinate	C03406	C10H18N4O6	290.123	-3.25	20 fmol	7.2	3.5	200 fmol	1.4	0.0
N-Acetyl-D-glucosamine	C00140	C8H15NO6	221.09	-2.6	200 fmol	3.2	1.0	2 pmol	1.5	0.0
N-Acetyl-L-aspartate	C01042	C6H9NO5	175.048	-0.79	2 fmol	5.3	2.3	2 pmol	1.4	-0.1
N-Acetylneuraminate	C00270	C11H19NO9	309.106	-2.78	2 fmol	4.9	2.1	2 pmol	1.5	0.0
N-Acetylputrescine	C02714	C6H14N2O	130.111	-0.84				2 pmol	1.7	0.2
N-Acetylserotonin	C00978	C12H14N2O2	218.106	0.98	200 fmol	1.7	0.0	200 fmol	3.8	1.5
N-Methylhistamine	C05127	C6H11N3	125.095	-0.57	20 pmol	5.7	2.6	20 pmol	2.0	0.4
N-Methylserotonin	C06212	C11H14N2O	190.111	1.55	20 pmol	2.0	0.3	200 fmol	3.4	1.2
N-Methyltryptamine	C06213	C11H14N2	174.116	2.02	20 pmol	1.9	0.2	200 fmol	3.8	1.5
NADH	C00004	C21H29N7O14P2	665.125	-1.45	20 fmol	5.4	2.4	2 pmol	2.1	0.4
Noradrenaline	C00547	C8H11NO3	169.074	-1.4						
Normetanephrine	C05589	C9H13NO3	183.09	-0.71	2 pmol	2.0	0.2	2 pmol	1.7	0.1
Octanoyl-CoA	C01944	C29H50N7O17P3S	893.22	-3	2 pmol	4.9	2.1	2 pmol	6.2	3.1
Orotate	C00295	C5H4N2O4	156.017	-0.89	20 fmol	2.6	0.6	20 pmol	6.7	3.5
Palmitoyl-CoA	C00154	C37H66N7O17P3S	1005.34	2.35	200 fmol	4.5	1.8	20 pmol	9.1	5.0
Phenethylamine	C05332	C8H11N	121.089	1.41				200 fmol	3.4	1.3
Phenylacetaldehyde	C00601	C8H8O	120.058	1.75	2 pmol	1.6	0.0	0.2 fmol	7.6	4.0
Putrescine	C00134	C4H12N2	88.1	-0.98				20 pmol	4.2	1.8
Pyridoxal	C00250	C8H9NO3	167.058	0.02	20 fmol	1.8	0.1	200 fmol	2.1	0.4
Pyridoxal phosphate	C00018	C8H10NO6P	247.025	-0.55	200 fmol	6.2	2.9	200 fmol	2.3	0.5
Pyridoxamine	C00534	C8H12N2O2	168.09	-1.23	200 fmol	3.9	1.4	2 pmol	1.8	0.2
Pyruvate	C00022	C3H4O3	88.016	-0.38	20 pmol	2.4	0.5	20 pmol	1.5	0.0
quinone	C00472	C6H4O2	108.021	0.21				20 pmol	7.6	4.1
Raffinose	C00492	C18H32O16	504.169	-3.36	2 fmol	7.8	3.8	2 pmol	1.5	0.0
Ribitol	C00474	C5H12O5	152.069	-2.53	2 pmol	2.9	0.8			
Riboflavin	C00255	C17H20N4O6	376.138	-1.05	200 fmol	1.2	-0.2	200 fmol	1.4	-0.1
Ribose	C00121	C5H10O5	150.053	-2.65	20 pmol	2.0	0.3			
Ribose 1-phosphate	C00620	C5H11O8P	230.019	-2.04	2 pmol	6.4	3.0			
Ribose 5-phosphate	C00117	C5H11O8P	230.019	-2.07	2 pmol	6.5	3.1			
Ribulose 5-phosphate	C00199	C5H11O8P	230.019	-2.07	200 fmol	6.1	2.8	20 pmol	1.7	0.2
Selenate	C05697	H2SeO4	145.912							
Serine	C00740	C3H7NO3	105.043	-3.42	200 fmol	5.2	2.2			
Serine	C00065	C3H7NO3	105.043	-3.42	200 fmol	5.3	2.3	20 pmol	1.5	0.0
Sinapyl alcohol	C02325	C11H14O4	210.089	1.36	20 pmol	4.5	1.8			
Sorbitol	C00794	C6H14O6	182.079	-2.68	20 fmol	4.1	1.5	20 pmol	1.5	0.0
Sorbitol 6-phosphate	C01096	C6H15O9P	262.045	-2.32	200 fmol	6.9	3.3	2 pmol	1.4	-0.1
Spermidine	C00315	C7H19N3	145.158	-0.62						
Spermine	C00750	C10H26N4	202.216	-0.66						
Succinate	C00042	C4H6O4	118.027	-0.53	2 pmol	5.1	2.2			
Succinate semialdehyde	C00232	C4H6O3	102.032	-0.47	200 fmol	4.5	1.8			
Succinyl-CoA	C00091	C25H40N7O19P3S	867.131	-6.7	200 fmol	6.9	3.3	200 fmol	3.4	1.3
Sucrose	C00089	C12H22O11	342.116	-2.63	0.2 fmol	5.5	2.4	20 pmol	1.5	0.0
Sulfite	C00094	H2SO3	81.9725	-2.7	20 pmol	1.7	0.1			
Taurine	C00245	C2H7NO3S	125.015	-2.19	200 fmol	4.0	1.5	2 pmol	1.6	0.1
Taurocholate	C05122	C26H45NO7S	515.292	0.79	20 fmol	1.6	0.0	200 fmol	7.0	3.6
Tetradecanoyl-CoA	C02593	C35H62N7O17P3S	977.314	1.84	20 fmol	4.6	1.9	2 pmol	8.5	4.6

Metabolite Information					HILIC (-)			RP (+)		
Name	**KEGG ID**	**Formula**	**Mass**	**logP**	**LOD**	t_R	k	**LOD**	t_R	k
Tetrahydrofolate	C00101	C19H23N7O6	445.171	-1.45						
Thiamin diphosphate	C00068	C12H19N4O7P2S	425.045	-0.1	20 pmol	9.1	4.7	20 pmol	2.1	0.4
Thiamin monophosphate	C01081	C12H18N4O4PS	345.079	-1.68	20 pmol	7.8	3.9	20 pmol	1.7	0.1
Thiamine	C00378	C12H17N4OS	265.112	-2.11	2 pmol	3.6	1.2	200 fmol	3.3	1.2
Trehalose	C01083	C12H22O11	342.116	-2.98	0.2 fmol	6.3	2.9			
Trypanothione disulfide	C03170	C27H47N9O10S2	721.289		2 pmol	10.4	5.5	20 pmol	2.0	0.4
Tryptamine	C00398	C10H12N2	160.1	1.21						
Tryptophan	C00078	C11H12N2O2	204.09	-1.1	200 fmol	2.1	0.3	200 fmol	3.4	1.3
Tyramine	C00483	C8H11NO	137.084	-0.14	2 pmol	2.0	0.2	2 pmol	2.2	0.5
Tyrosine	C00082	C9H11NO3	181.074	-2.39	200 fmol	4.0	1.5	2 pmol	1.7	0.1
UDP-D-galactose	C00052	C15H24N2O17P2	566.055	-6.5	2 fmol	6.4	3.0	2 pmol	1.4	-0.1
UDP-glucuronate	C00167	C15H22N2O18P2	580.034	-1.21	2 fmol	7.1	3.4	2 pmol	1.4	-0.1
Urate	C00366	C5H4N4O3	168.028	-1.12				0.2 fmol	7.6	4.0
Urea	C00086	CH4N2O	60.0324	-1.78						
Xanthine	C00385	C5H4N4O2	152.033	-0.65	2 pmol	4.6	1.9			
Xylitol	C00379	C5H12O5	152.069	-3.2	2 pmol	3.6	1.3			
Xylose	C00181	C5H10O5	150.053	-2.57						

2 Improving the detection of low abundant metabolites by combining ion intensities of multiple LC/MS runs

David Jonathan Fischer, Christian Panse and Endre Laczko

Functional Genomics Center Zürich, University of Zürich and ETH Zürich, 8057 Zürich, Switzerland

D.J. Fischer contributed to the design of the experiments, analysed the data, developed the algorithm and wrote the manuscript. C. Panse was involved in the implementing of the algorithms into a R package. E. Laczko supervised the study and contributed to the experimental design.

2.1 Abstract

Raw data processing, i.e. the extraction of compound identities and quantities, is a very important task in untargeted metabolomics. Coupled to high accuracy liquid chromatography mass spectrometry (LCMS) data, a vast amount of data is accumulated and file sizes approaching gigabyte dimensions are typically achieved with current instruments. In order to compare different LCMS experiments, a data reduction workflow to reduce complexity is essential. The aim of data processing typically involves the retrieval of mass-to-charge, retention time and ion intensity peaks. Current data processing tools focus on a peak detection strategy where initially each LCMS run is treated independently from each other. Subsequent data processing for the alignment of different samples is then calculated on reduced peak tables. In this work we describe a novel approach that involves the merging of all LCMS datasets of a given experiment as a first step of raw data processing. The merged LCMS dataset is reduced in size because it contains an overlay and the sum of ion intensities from all LCMS runs. Peak detection is then performed only on this merged dataset and quantification of the signals in each sample is guided by the peak location in the merged data. We compared our data processing strategy with algorithms of the Bioconductor package XCMS and demonstrate that a large amount of low abundant signals can only be detected from the merged data while they are missed during peak detection in single LCMS runs.

2.2 Introduction

In a untargeted LCMS based metabolomics experiment, raw data processing is the first fundamental part in the computational pipeline and a couple of software tools have been developed for this purpose (Sugimoto et al., 2012). The core function of LCMS raw data processing is to identify and quantify as much metabolites as possible in a series of data sets. This task typically consists of three steps: 1) Peak detection in a 3-dimensional space: mass-to-charge ratio (m/z), retention time (RT) and ion intensity. 2) Linear or non-linear transformation of the retention time. This step is important because analytical variations in the chromatographic dimension usually cause shifts in the elution profile. 3) Alignment of the identified features (m/z and RT location) across different LCMS runs. As a result, the algorithm produces a

data matrix that can be used for statistical analyses and biological interpretation. A good raw data processing algorithm should have the following characteristics: 1) It is sensitive enough to detect even the smallest peaks and 2) Is robust against false detection of noise (Tautenhahn et al., 2008).

A critical point in LCMS raw data processing is the detection of low abundant signals. Typically, the dynamic range of ion intensities is quite broad and low abundant features are characterised by a low signal-to-noise ratio (SNR). In most metabolomics experiments, a multitude of LCMS runs from a similar biological matrix is analysed. It is expected that each metabolite signal is always located at the same position, i.e. has always the same RT and m/z value. We therefore hypothesized that overlaying of the raw data signals from multiple LCMS runs leads to a better SNR and eventually might improve the detection of low abundant metabolites.

2.3 Experimental Section

Colorectal biopsy samples were collected from surgically removed colon. The biopsies with fresh weights between 3 and 5 mg were snap frozen in liquid nitrogen and stored at -80°C until use. 100 µL methanol per 3 mg biopsy was added and the tissue was disrupted using a glass homogenizer. Debris was removed and the extract was diluted 1:5 with 5 mM ammonium formiate for reversed phase LCMS. For HILIC LCMS, the extract was dried under vacuum and resuspended with injection solution prior to LCMS analysis (90% acetonitrile, 9% methanol, 1% water, 50 mM ammonium acetate at pH 9). For the reversed phase method a 50 mm self packed capillary column with a internal diameter of 200 µm was used. The chromatography material HSS T3 (Waters Corp., Miford USA) was filled into the capillary. We used a 10 min gradient from 5 mM ammonium acetate to 90% isopropanol/10% acetonitrile with a flow rate decreasing linearly from 5 to 3 µl/min. For the HILIC method we used also a self packed capillary column with the following properties: 15 mm column length, BEH amide material (Waters Corp.), 200 µm inner diameter. The elution buffers consisted of A: water and B: 95% acetonitrile and both buffers contained 0.5 mM ammonium acetate at pH 9. The gradient was 90% B to 50% B in 10 minutes using a linearly decreasing flow rate from 6

to 4 µl/min. For the analysis of the biopsy samples, we collected 30 normal mucosa or cancer biopsies. These samples were analysed in triplicates and additionally a pool sample from the 30 extracts was injected 10 times. In total, a set of 100 LCMS runs was created. The MS data was acquired on a Synapt G2 in MSE negative (HILIC) or positive (reversed phase) mode. The low collision energy full scan channel was converted to centroid and vendor independent netCDF format using MassLynx 4.1 and the integrated conversion software DataBridge (Waters). To compare our algorithms with xcms algorithms, netCDF data were processed with xcms (Smith et al., 2006) using the matchedFilter algorithm (fwhm = 4, snthresh = 5, step = 0.02, mzdiff = 0, max = 50) or the centWave algorithm (ppm = 25, snthresh = 5, peakwidth=c(1,60)). The retention time of the peaks were aligned using the obi-warp algoritm with parameters profStep=1 and distFunc="cor". Peak tables were grouped using the parameters mzwid = 0.01 and bw = 2. To annotate C13 isotopes we used the CAMERA package (Kuhl et al., 2012). For this purpose, the functions “groupFWHM” and “findIsotopes” were used. To evaluate the algorithm performance with differently sized sets of LCMS runs, the complete workflow (xcms and CAMERA) was applied to 2, 5, 10, 20, 50 or 100 samples. The selection was made randomly with the exception of one pool sample which was included every time. This pool sample was used as the reference for retention time alignment.

2.4 Theory

Full scan high accuracy mass spectrometry, coupled to chromatography are powerful and popular methods to profile hundreds to thousands of metabolites in biological samples. Such an untargeted experimental strategy requires algorithms capable to reliably identify and quantify a large amount of signals. Typically, a single full scan LCMS metabolomics dataset consists of hundred thousands data points and data sizes around 100 MB or more. For the analysis of multiple metabolomics datasets such as comparison of different biological conditions there is a need of data reduction in order to process several to hundreds of LCMS datasets. A common strategy to reduce required memory space in full scan LCMS data processing is to generate retention time/mass-to-charge peak lists as a first step (Sugimoto et al., 2012). In order to detect signals with a very low signal intensity and poor signal-to-

noise ratio (SNR) the algorithm paramaters are defined very generously, i.e. the SNR threshold is set to a very low value. The disadvantage of this approach is to risk the detection of a high amount of false positive peaks. The basic idea behind the here presented study is based on the fact that LCMS datasets are fairly well reproducible in the mz and retention time locations of every analyte. While low abundant mass signals with low SNR values in a single LCMS run are difficult to be reliably detected, a combined analysis of all signals from multiple LCMS runs might decrease the amount of random noise and improve ion statistics (Figure 2.1). An algorithm with the ability to detect peaks on a dataset which combines ions of multiple LCMS runs could therefore potentially detect low abundant signals with a higher robustness compared to algorithms where peaks are detected on single LCMS runs alone. Here we will describe our raw data processing approach. It aims for the combination of the mass and retention time dimension of multiple LCMS runs. Using this approach we generate overlays of the mass or retention time dimension of all samples. The peak detection process (mass peaks or retention time peaks) is carried out on these overlays. All described algorithms were developed in the open source environment R (R Development Core Team, 2013) or modified using existing R and Bioconductor packages (Du et al., 2006; Smith et al., 2006; Tautenhahn et al., 2008). We implemented the presented workflows in a new open source R package called *cosmiq (**Co**mbine **s**ingle **m**asses **i**nto **q**uantities)*. The package was designed to support the xcmsSet class and can be therefore used together with other functions built around the xcms package. The five steps of the processing strategy of cosmiq are shown in Figure 2.2.

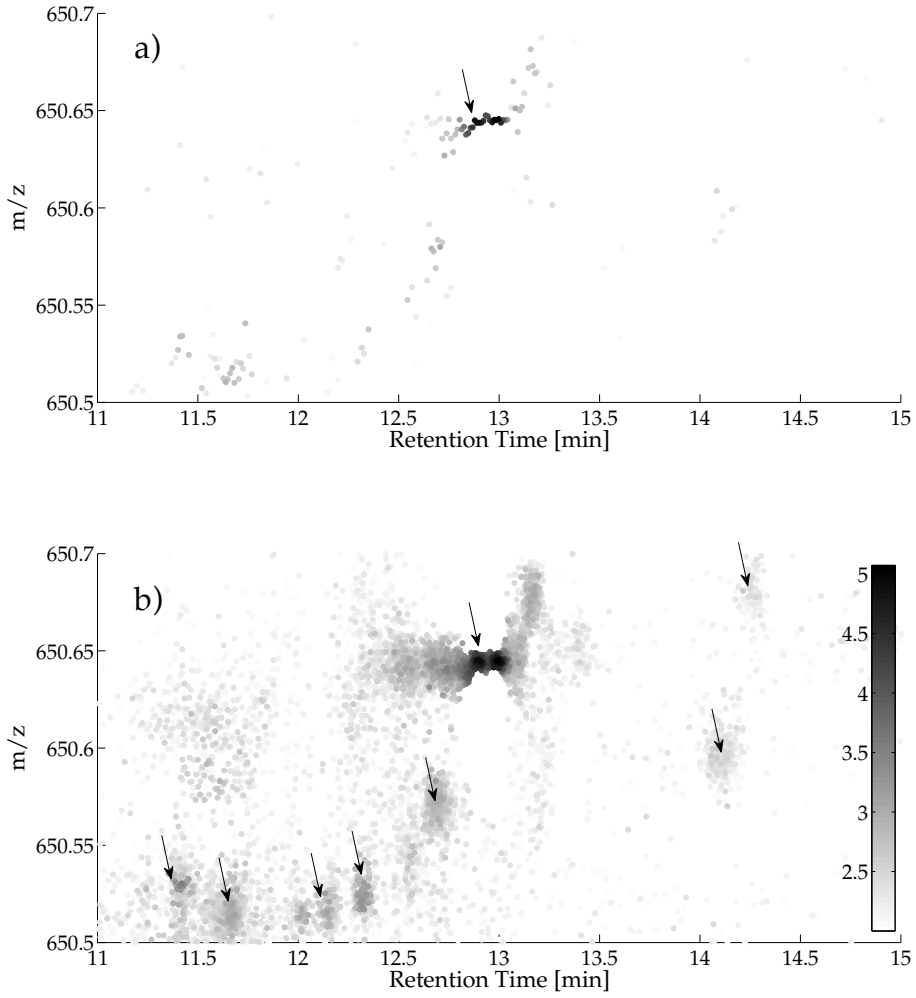


Figure 2.1: Example region of metabolomics raw data based on high accuracy mass LCMS. The masses range between 650.5 and 650.7 Da is shown. Data were recorded in centroid mode. Centroids of a single (a) or an overlay of the centroids of 102 LCMS runs (b) are shown. The color indicates the centroid intensity and the colorbar is displayed along a log(10) scale. The retention time of the 102 LCMS runs in b) were aligned before data were plotted. Arrows indicate dense regions of centroids where a mz/RT peak is clearly visible by visual inspection.

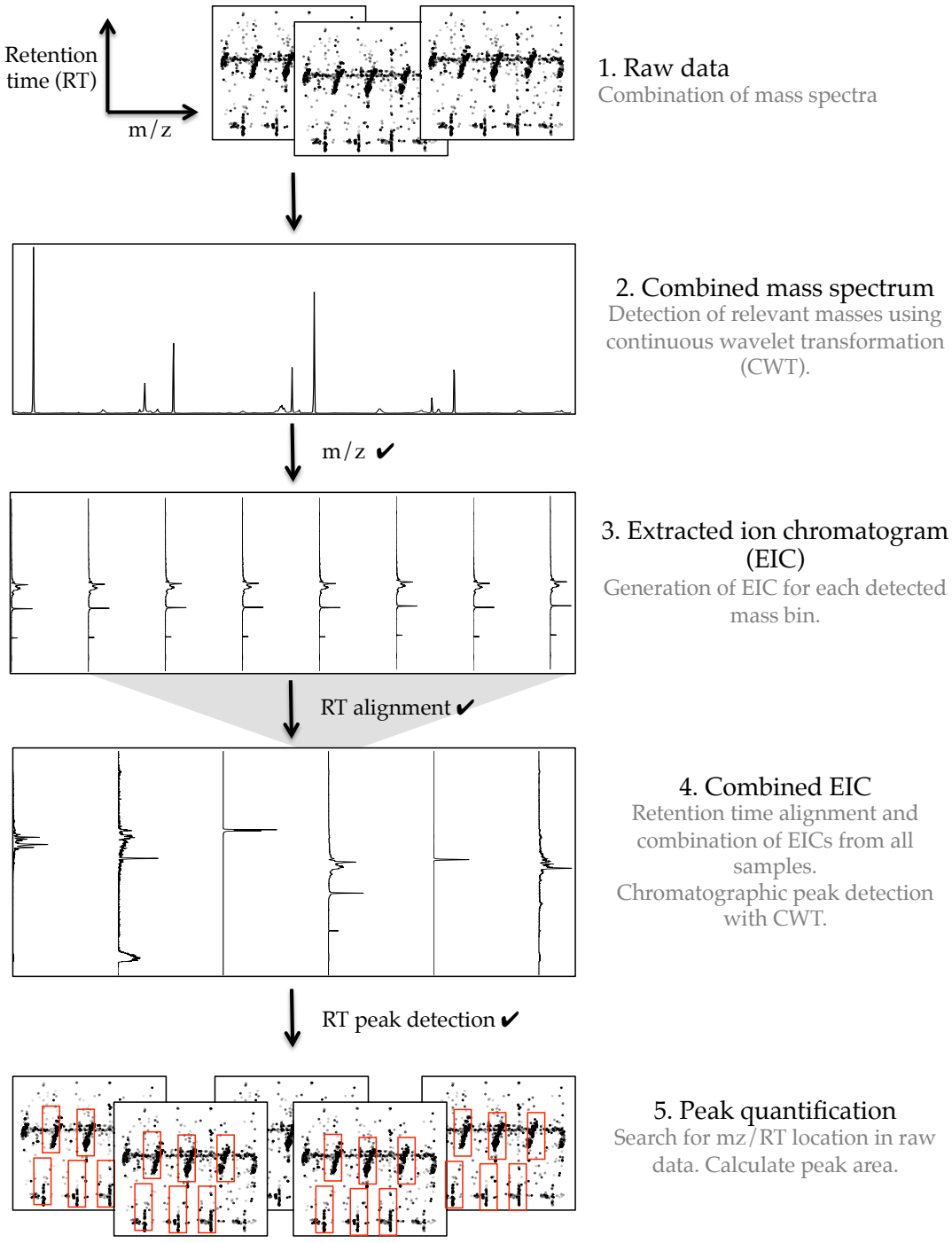


Figure 2.2: The raw data processing pipeline of cosmiq.

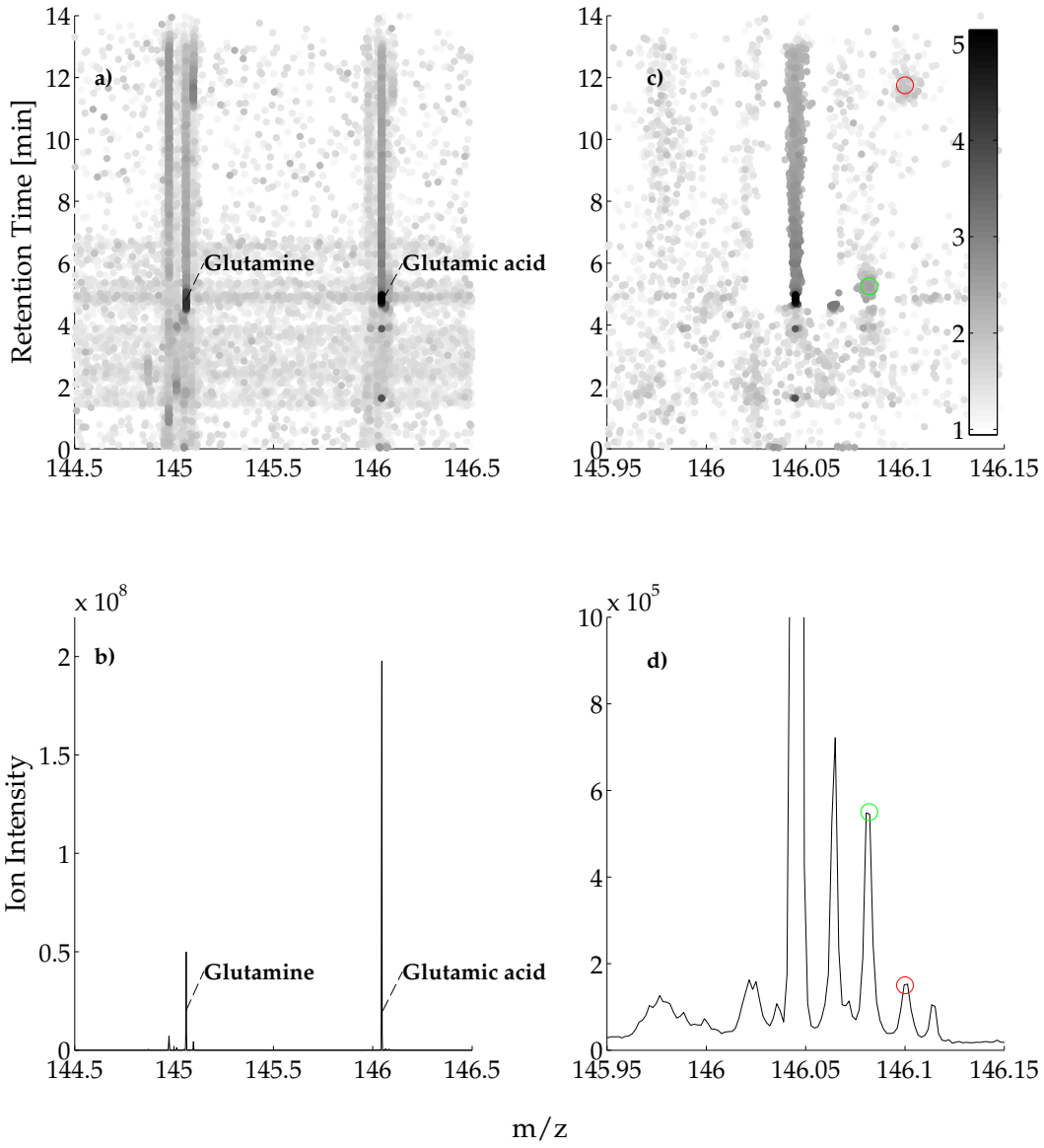


Figure 2.3: Step 1: Combination of mass spectra for the detection of relevant mass bins. Raw data of a single LCMS run is shown around the region of the metabolites glutamine and glutamic acid (a) and a close-up of the region around glutamic acid (c). The combined mass spectrum of the same sample and the regions as seen in a) and c) is shown in b) and d), respectively. The intensity of the raw data centroids is indicated by a grayscale color. The colorbar represents log(10) transformed intensities.

2.4.1 Step 1: Combination of mass spectra

The first two processing steps search for relevant mass bins in the dataset. Binning of the mass dimension was criticised because it is difficult to select an optimal bin size without a priori information of occurring masses (Aberg et al., 2008). In order to select for optimal bins, we first calculate a combined spectrum. This approach of overlaying and summing intensities of single scans together is prominent for applications in flow injection mass spectrometry and leads to improved ion statistics (Fuhrer et al., 2011). Mass spectra not only from all scans of a single LCMS run alone are combined but

from all acquired datasets. As a result, signal to noise ratio increases for each additional LCMS run (Figure 2.3). For example, high abundant metabolites with intensities around 10000 counts like glutamine and glutamic acid are clearly visible in the raw centroid data as well as in the combined mass spectrum (Fig 2.3 b). Low abundant masses around 100 counts are difficult to spot in the raw LCMS data but show a clear peak in the combined mass spectrum (red and green circle, Figure 2.3 d).

2.4.2 Step 2: Detection of relevant masses

Based on this combined mass spectrum we then determine location and boundary of each mass. A peak detection algorithm based on continuous wavelet transformation (CWT) is used for this step (modified from Du et al., 2006). Peak detection based on CWT has the advantage that a sliding scale of wavelets instead of a single filter function with fixed wavelength is used (Tautenhahn, 2009). This allows for a flexible and automatic approximation of the peak width. As a result it is possible to locate both narrow and broad peaks within a given dynamic range. To detect the optimal peak locations, regional maxima on each scale of the CWT transformed data are calculated. The regional maxima on the scale with the highest intensity are selected as optimal peak locations (as described in Du et al., 2006). The adjacent local minima on the selected scale maximum are a good estimator of the real peak width (Tautenhahn, 2009). However, regional maxima in the CWT space are only a reliable indicator for a good peak location if there are no overlapping peaks in the dataset (Tautenhahn et al., 2008). The scale with the highest intensity detects all overlapping peaks as a single one (Figure 2.4 b). In order to test if there exist two or more overlapping peaks we examine the raw data at each position where an optimal peak was located by the regional CWT maximum. The width of the suspected original peak at a selected intensity treshhold (e.g. width at half peak maximum) is investigated. If there exist overlapping peaks, the width of the peak is divided. The new optimal CWT scale is then estimated as the width of the broadest overlapping peak (Figure 2.4 d).

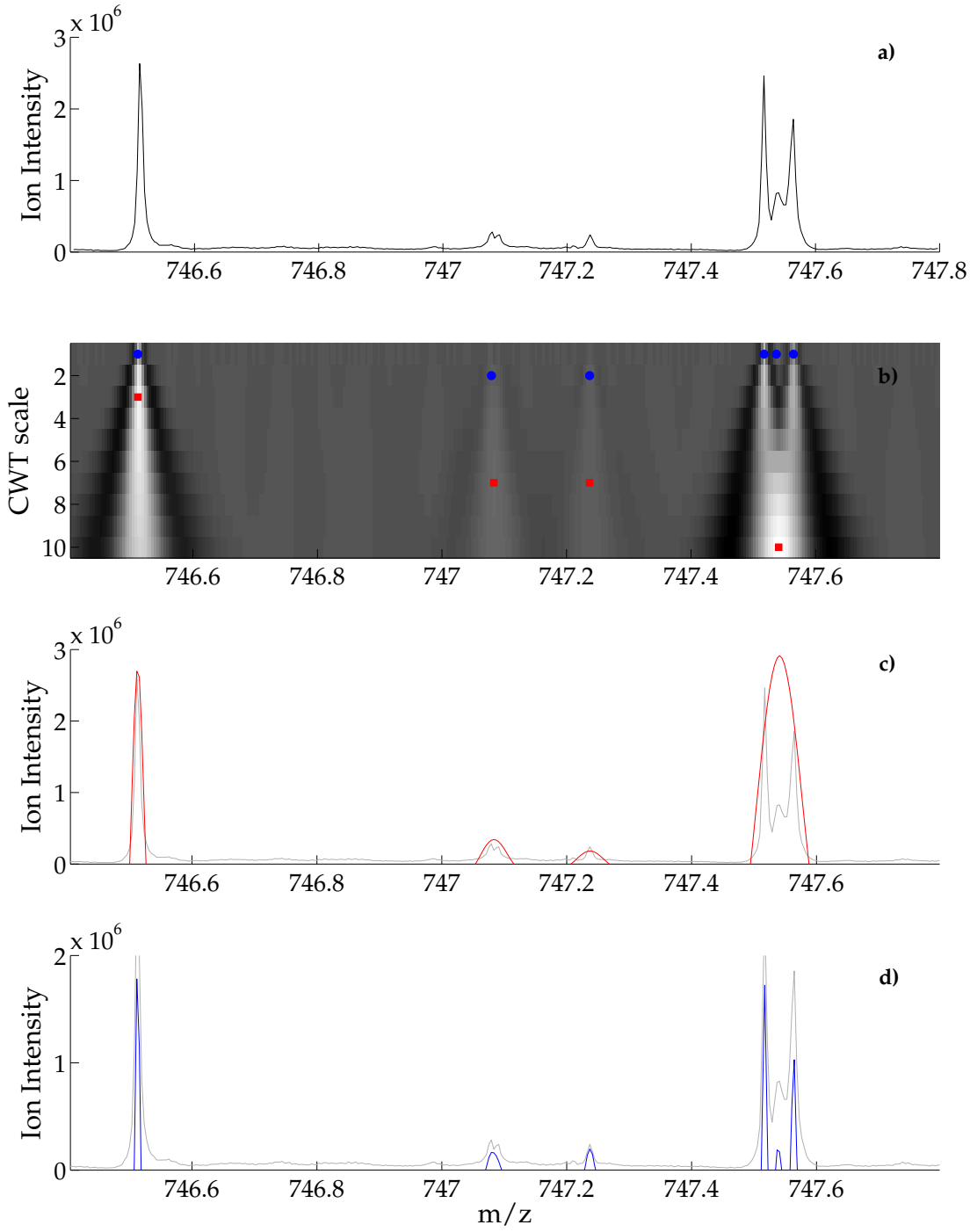


Figure 2.4: Step 2 and Step 4: Feature detection using continuous wavelet transformation (CWT). Close-up of a combined mass spectrum between 746.5 and 747.8 Da (a). This spectrum was transformed with scales between 1 and 10 using CWT (b). Peak locations as indicated by local CWT scale maximum are shown as red dots. Local scale maxima overlaid to the raw data (c, red curves). The overlapping peaks between mz 747.5 and 747.6 are only resolved if the optimal scale in the region is approximated using the raw data (blue curves in d), blue dots in b), see text for details).

2.4.3 Step 3: Generation and combination of extracted ion chromatograms

Until now only the mz information was considered. In the following processing steps the chromatographic information will be added. For the comparison of different LCMS datasets it is important to consider RT shifts.

These shifts are typically caused by technical variations and need to be corrected before chromatographic peaks between different LCMS runs are aligned (Smith et al., 2013). For this purpose cosmiq implements the warping algorithm OBI-WARP which is included in the xcms package and can be directly applied on the raw data (Prince and Marcotte, 2006). For each detected mass in step 2 we calculate an extracted ion chromatogram (EIC). In order to determine the elution time for each detected mass, the EICs of every mass are combined between all acquired runs. Again, this combination approach aims for an improvement of the signal-to-noise ratio (SNR). As an example we demonstrate the detection of four hexose-phosphate isomers with the monoisotopic mass 259.0223 ([M-H] adduct, Figure 2.5). Before RT alignment, the EICs are highly dispersed along the RT axis (Figure 2.5a). After RT correction, the EICs between all the acquired LCMS runs are well aligned (Figure 2.5b). The monoisotopic peaks of the four isomers can be seen in one single EIC, but the signal is noisy (Figure 2.5c). The SNR of these four chromatographic peaks becomes better when all aligned EICs from Figure 2.5b are combined (Figure 2.5e). This effect becomes more evident if the less abundant isotopic mass with one C(13) of the same molecule is investigated: The four isomer peaks are hardly distinguishable from the noise in a single LCMS run (Figure 2.5d) but become clearly visible in the combined EIC of all LCMS runs (Figure 2.5f).

2.4.4 Step 4: Detection of chromatographic peaks

Based on the combined EICs there is another peak detection step to be performed. The algorithm as described for the peak picking of m/z signals in Step 2 is used also for peak picking in the retention time domain. The final result is a peak table with location and boundaries of each mz/RT feature. This information will be further used to locate the relevant position in every single LCMS dataset in order to quantify sample specific feature intensities. Because the mz/RT features were detected on the combined mass spectra or EICs of all samples it is not necessary to align features between different LCMS runs as for the typical raw data processing workflow (Sugimoto et al., 2012). Instead, a data matrix with intensity values for every mz/RT feature and every sample can be immediately calculated.

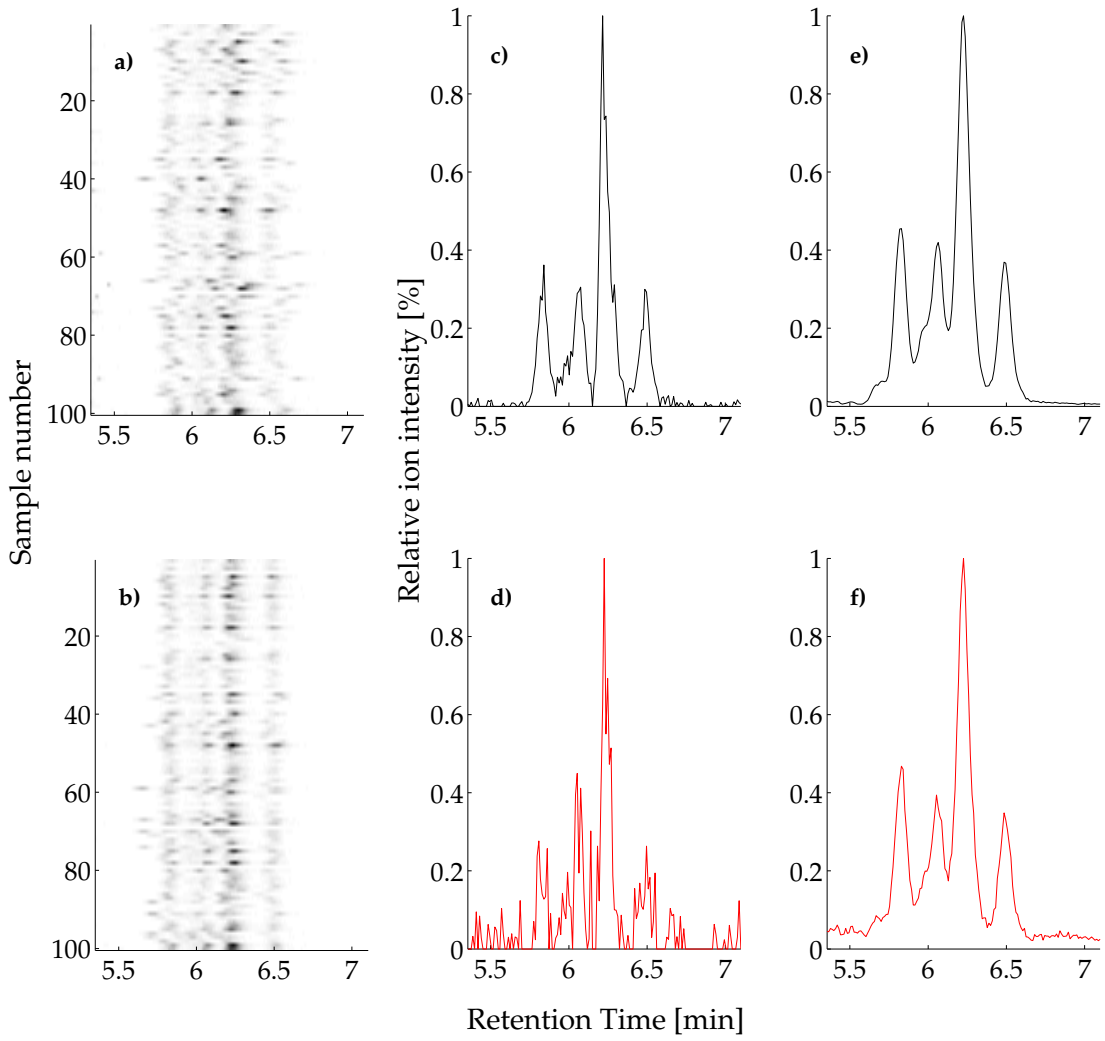


Figure 2.5: Step 4: Chromatographic peak detection using continuous wavelet transformation. Extracted ion chromatograms (EIC) are shown for 100 LCMS samples before (a) or after retention time alignment (b) Ion intensity for each sample is shown in grayscale intensity. c) EIC of a single LCMS sample showing the monoisotopic mass of hexose-phosphate (m/z 259.0223). e) Peak detection of the same mass and retention time window, but with a combined EIC of 100 LCMS runs. The EIC for the same molecule containing one C(13) isotope is show in d) for a single EIC and for the combined EIC in f).

2.4.5 Step 5: Localisation and quantification of detected peaks

With the information about their position in the combined datasets, each individual mz/RT feature is then located in the raw data. Due to the retention time correction, each feature is expected at the same RT position as in the combined EIC. However small shifts in retention time still occur for most of the peaks (See for example the EIC from sample 59 in Figure 1.5 b). In order to locate the correct position of each feature, the EIC of the selected mass is calculated for the whole retention time. This EIC is filtered with CWT using only the scale where the feature was optimally located on the combined EIC in step 3. Local maxima are calculated on this transformed data and the

maximum with the closest position to the expected retention time is chosen. Local minima to the maximum of the transformed data are selected as peak boundaries as suggested by (Tautenhahn, 2009) and the peak area is then calculated on the raw EIC.

2.5 Results and Discussion

2.5.1 Evaluation of different algorithms

A peak list with intensity values for mz/RT features of every sample is produced as the output of the five data processing steps. A further step to deisotope peaks or annotate adducts is not included. As already mentioned, the cosmiq package is intended to provide alternative raw data preprocessing within a metabolomics workflow using R. Thus, other xcms based functions or packages can be used in combination with the processing pipeline of cosmiq. To evaluate our algorithms, we designed an experiment with the aim to test the significance of analysing multiple LCMS runs together. The main idea to develop cosmiq was to improve the detection of low abundant metabolites by combining the ion intensities of multiple LCMS runs before peak detection is performed. Our central hypothesis states that the more LCMS runs are added to the analysis, the more robust will be the detection of low abundant signals. We created two different metabolomics datasets for this testing purpose: One dataset using hydrophilic interaction chromatography (HILIC) for a polar metabolome and reversed phase chromatography (RP) for a lipidome analysis. Each dataset consists of capillary flow based LCMS analyses of colorectal tissue metabolite extracts. 30 normal mucosa and cancer biopsies were collected and analysed in three technical replicates. Additionally, these 30 samples were collected together in a pool sample. This pool sample was injected 10 times and in total both datasets consist of 100 LCMS runs, each containing the analysis of a similar biological matrix. From these 100 LCMS runs we randomly selected 2, 5, 10, 20, 50 or 100 for a joint data analysis. The LCMS analyses were done in full scan mode to allow for a non-targeted data acquisition. We compared the cosmiq workflow with two peak picking algorithms from the xcms package: *matchedFilter* and *centWave* (Smith et al., 2006; Tautenhahn et al., 2008). matchedFilter is the original algorithm of xcms and centWave was developed for high resolution centroid LCMS data. Both algorithms were tuned in a way

that low abundant signals with a poor SNR are considered (SNR threshold 5). Other peak detection parameters were tuned in order to reveal an equal amount of features for the analysis of five samples. The cosmiq parameters were tuned with a low SNR threshold (5) for the mass detection and a moderate SNR threshold (20) for the chromatographic peak detection step. The retention time parameters of the obtained peak lists were aligned using the obiwarp function. As the retention time alignment with cosmiq is also based on this algorithm, the time alignment step is identical in all three workflows. The final grouped peak list was used as basis for the evaluation process.

2.5.2 Effectiveness of the different algorithms

For a non-targeted metabolomics workflow it is difficult to assess the quality of raw data processing. Typical approaches use known metabolite standards as references (Nordström et al., 2006; Wei et al., 2011). Such an approach however is restricted to the interpretation of a small set of spiked analytes and not suitable to estimate the quality of thousands of annotated features. Other feature evaluation studies aimed for the calculation of precision and recall values as a information retrieval measure (Tautenhahn et al., 2008; Kenar et al., 2014). For these approaches it is necessary to define a robust ground truth set of features, which has to be recognized by the algorithms. One of these studies generated an artificial LCMS dataset using a simulator software, where the ground truth was retrieved from the simulator (Kenar et al., 2014). In another study the authors calculated a ground truth dataset based on an intersection peak list from different algorithms (Tautenhahn et al., 2008). For the evaluation of the feature detection quality of our dataset with 100 samples, we included the package *CAMERA* to our workflow (Kuhl et al., 2012). The main function of CAMERA is to deconvolve the feature list output of xcms and to annotate C(13) isotopes and adducts. Our assumption was that a good metabolomics data processing workflow reliably detects not only single masses, but identifies all connected isotope patterns. We therefore calculated the number of features which could be linked to an isotope pattern and compared it to the total number of detected features. The resulting ratio between 0 (no peak could be assigned to an isotope pattern) and 1 (all peaks were annotated within an isotope pattern) was used as quality measure. This

approach has several advantages: 1) Every detected feature of the non-targeted data processing is considered and therefore all low abundant signals. 2) Real metabolomics data and no simulation data is used. With this measure we observed that the quality of the feature detection is independent of the amount of LCMS samples in the data processing pipeline (Figure 2.6 a and b). For both, the HILIC and the RP datasets we observe a relatively constant ratio of annotated isotopes compared to unannotated features by CAMERA. This result suggests that the detection of false positive features is independent of the amount of LCMS runs. For the HILIC measurement the ratio of isotope pattern/all peaks was between 0.3 and 0.35 for the cosmiq algorithm with a slight increase towards multiple LCMS runs. For both the matchedFilter and centWave it was around 0.1. For the RP dataset the ratio was around 0.45 for cosmiq, 0.25 for centWave and 0.2 for matchedFilter. Although this result suggests that cosmiq provides a better peak detection effectiveness compared to the xcms algorithms, we want to point out that the selection of data processing parameters were chosen quite unrestrictive in terms of SNR parameters. Therefore a high number of unspecific signals are detected and a low amount of isotope patterns has to be expected. The total amount of detected monoisotopic masses ([M]) was quite constant in the HILIC dataset when matchedFilter or centWave was used (Figure 2.6 c). In the RP dataset, the total number of [M] slightly increased from 2 to 10 LCMS runs but remained constant between 10 and 100 runs (Figure 2.6 d). Interestingly, the number of LCMS runs had an enormous impact on the amount of detected [M] when cosmiq was used for both HILIC and RP (Figure 2.6 c/d). For a small amount of LCMS runs (2-5), the number of [M] is minimally higher in the cosmiq dataset versus the xcms algorithms. However this number drastically increases almost ten-fold for the HILIC dataset and 4-fold for the RP dataset when 100 LCMS runs are processed. In summary, the addition of LCMS runs to the processing pipeline does not infer with the overall peak detection quality for none of the three algorithms. However, for the data processing with cosmiq the amount of detected monoisotopic peaks and therefore the amount of putative metabolites increases with the number of LCMS runs whereas the xcms algorithms always detects a constant amount of monoisotopic peaks irrespective of the number of LCMS runs. This result

indicates that the combination of LCMS runs with cosmiq can improve raw data processing in terms of an increased metabolite detection rate.

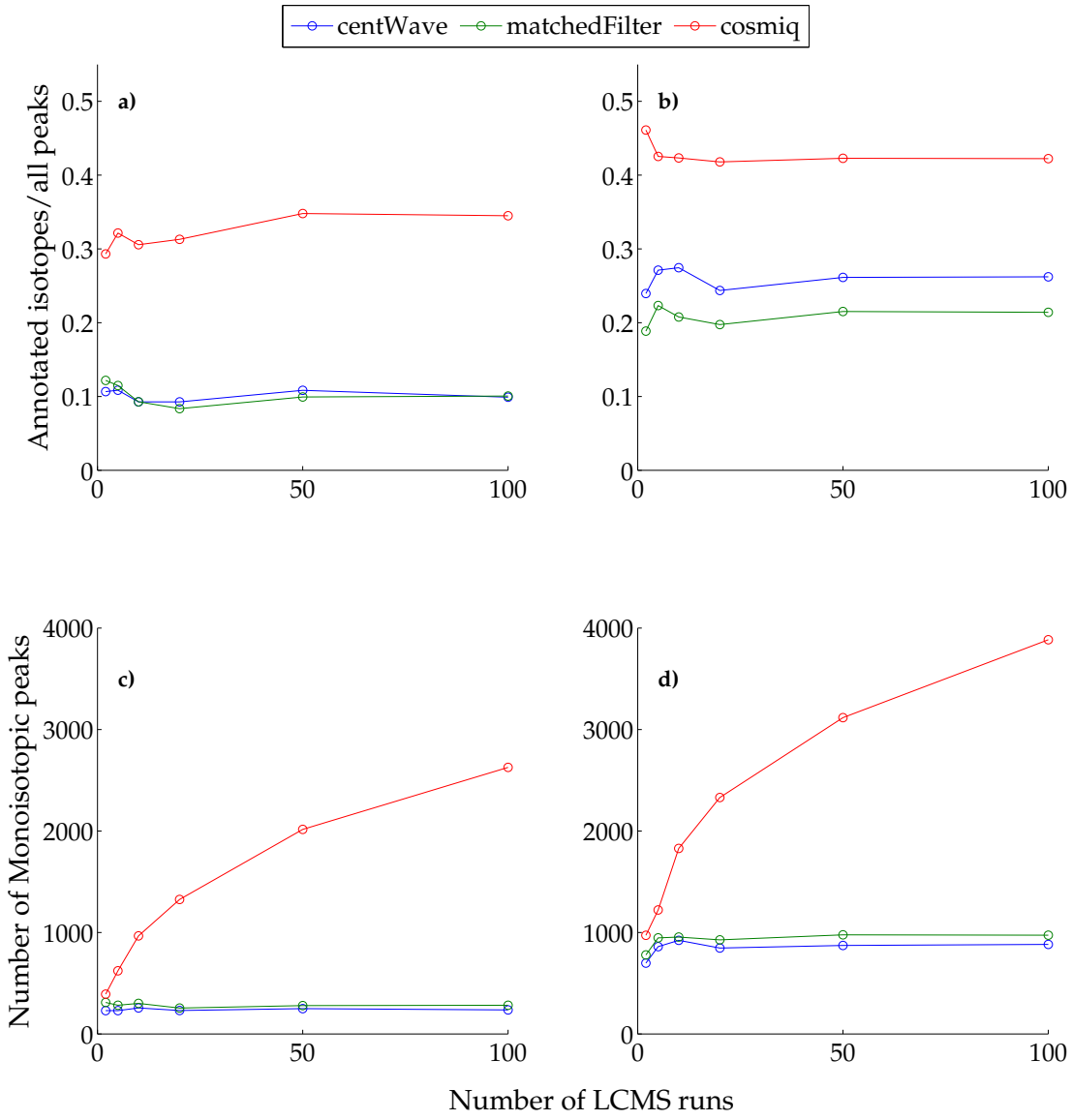


Figure 2.6: Number of detected features and comparison of three different peak detection algorithms. Peak detection quality was calculated as the ratio of annotated isotopes/all peaks shown for a HILIC (a) and a RP dataset (b). Number of detected monoisotopic masses as calculated by the package CAMERA is shown for HILIC (c) and RP (d).

2.5.3 Dynamic range of the detected metabolome

An increased number of metabolite detection by cosmiq compared to the xcms algorithms can be explained by an improved detection of low abundant compounds. As illustrated in Figure 2.5 the combination of LCMS samples improves the signal to noise ratio of low abundant signals. Consequently, the metabolites which are missed by the xcms algorithms due to a poor SNR in

each sample, might be detected by cosmiq. In order to test this hypothesis, we investigated the dynamic range of the metabolome in each of our datasets. We matched all found monoisotopic masses to the human metabolome database (Wishart et al., 2013) and annotated metabolites within a mass accuracy of 0.01 Da. The annotated metabolites were classified in three groups: 1) Detected by all algorithms, 2) detected by matchedFilter, centWave or both, 3) detected only by cosmiq. In order to investigate the effect of multiple LCMS runs on the annotation, we compared the lowest number (2) and the highest number of joined LCMS samples (100).

For each of the annotated metabolites, we calculated the mean intensity values across every sample. For the metabolites detected by all algorithms there was a slight increase of annotated metabolites from 2 to 100 joined LCMS samples (Figure 2.7). For the HILIC dataset, 41 metabolites were detected in 2 and 49 metabolites in 100 LCMS runs, respectively. For the RP dataset there was an increase from 141 to 213 detected metabolites. Conversely, there was a slight decrease of annotated metabolites which were unique for the matchedFilter or centWave dataset: 91 to 84 for the HILIC dataset and 146 to 135 for the RP dataset, respectively. For the features which were uniquely detected with cosmiq there was a more than 4-fold increase of detected metabolites for both HILIC and RP. These numbers indicate again that multiple joined LCMS samples lead to an increased detection of metabolites in the data processing pipeline of cosmiq. Remarkably, not only the number of detected metabolites increases but also the dynamic range of intensities decreases for the metabolites which were only detected by cosmiq. If only 2 LCMS runs are considered, the interquartile range of peak intensities lies between 619 and 2774 for the HILIC dataset and between 1788 and 9812 for the RP dataset. For 100 LCMS runs however, the interquartile range drops to 106/794 for HILIC and to 363/1326 for the RP data. The peak intensity range of solely cosmiq specific metabolites is therefore up to ten times lower compared to the matchedFilter or centWave specific metabolites. Altogether these numbers support our conclusion that our raw data processing approach where multiple LCMS runs are combined before feature detection leads to a better detection of low abundant metabolites. With the joined analysis of 100 LCMS samples we achieve up to a ten times increased feature detection rate.

A substantial amount of these features have very low mean intensity values below 1000 and even below 100 ion counts.

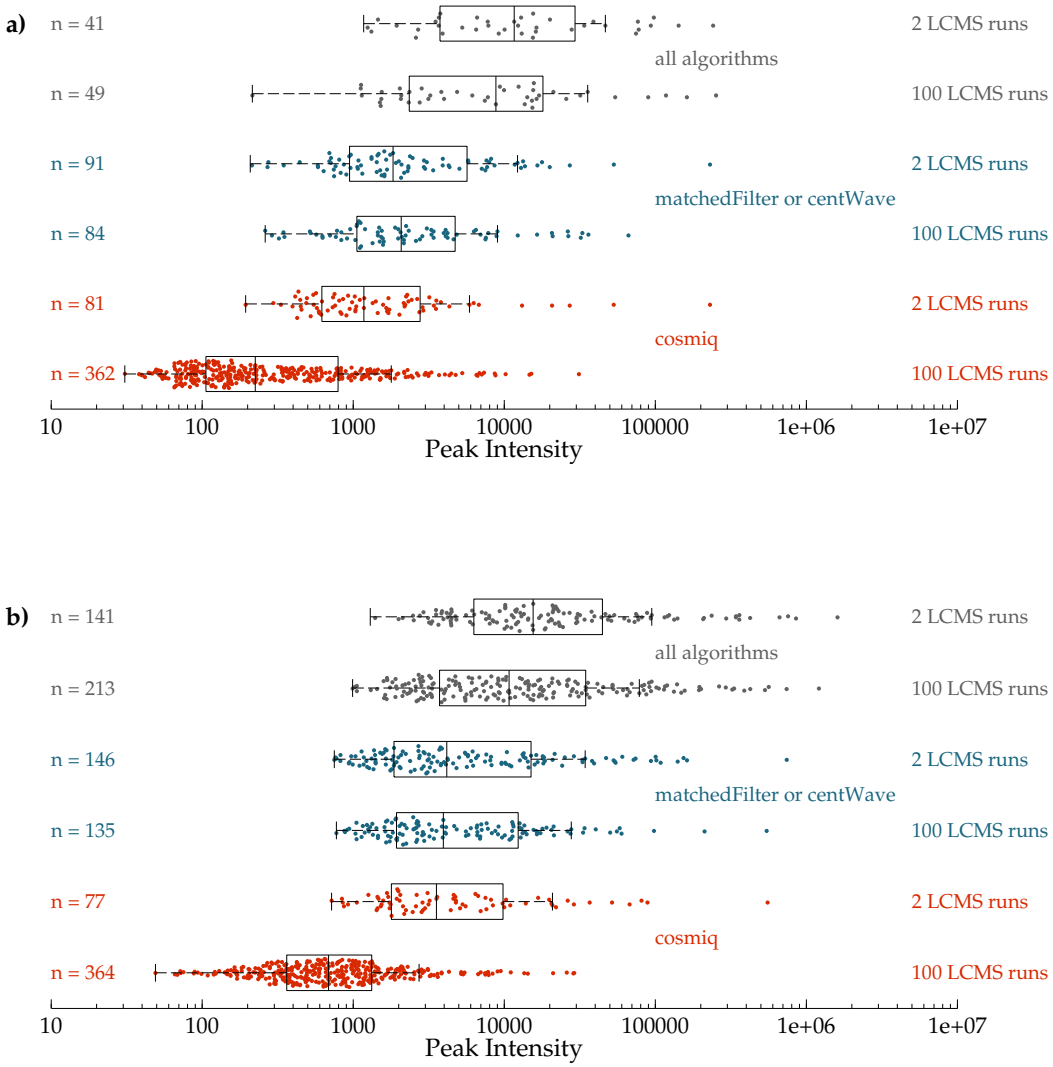


Figure 2.7: Comparison of the dynamic range of metabolome intensities. Monoisotopic masses were annotated to the human metabolome database and peak intensities of the metabolites are shown for the HILIC (a) or RP dataset (b). Metabolites which were found with all algorithms are shown in gray. Metabolites identified by the matchedFilter, centWave or both of these algorithms are blue. Metabolites only annotated after processing with cosmiq are shown in red.

2.5.4 Computation time

Especially the feature detection steps using continuous wavelet transformation require much computation time in cosmiq. However, since the peak detection is performed on the combined data of all LCMS runs, this step is only computed once for all LCMS samples. As a result, the average computing time for each LCMS run decreases if multiple samples are included to the analysis (Figure 2.7). Since the amount of detected features

varied for each algorithm, we considered the computation time as a function of the feature number. For only two LCMS runs the computation time of cosmiq is almost one order of magnitude slower compared to the centWave algorithm, which also outperformed the matchedFilter. However for 50 or more LCMS runs, cosmiq performs equally fast. The runtime was calculated wall-clock timewise on a Intel Core i7 2.6 GHz processor with 16 GB RAM and Mac OSX 10.9 as the operating system.

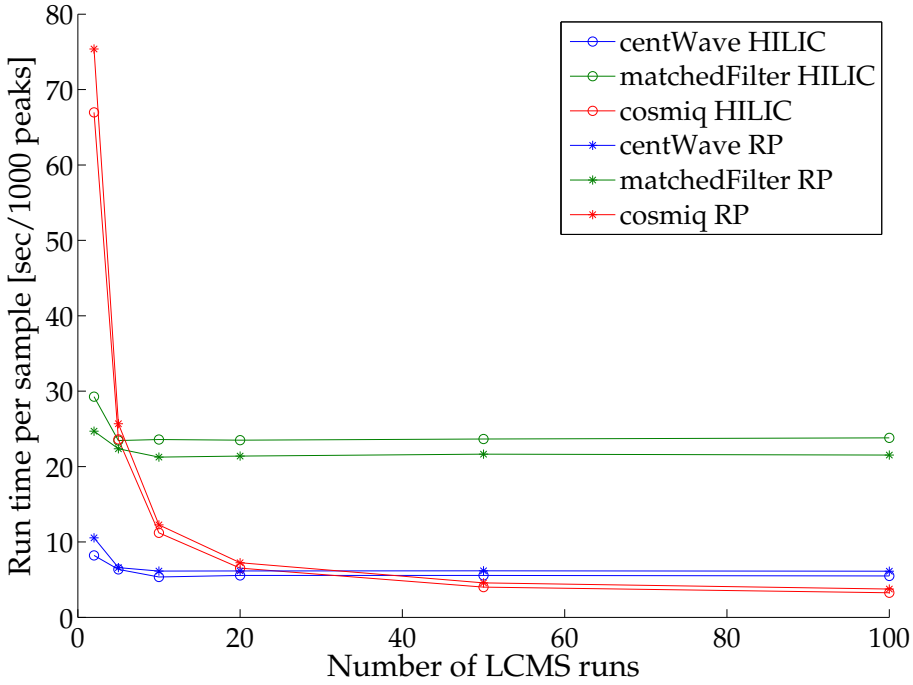


Figure 2.8: Computation time of the three peak detection algorithms centWave, matchedFilter and cosmiq. The total runtime of each processing step was averaged across each sample and normalized to the total number of detected peaks.

2.6 Conclusion

We present a novel approach for the processing of raw LCMS data with a focus on metabolomics or lipidomics applications. It is based on the idea that an overlay of multiple raw LCMS runs increases ion statistics and therefore the SNR. We integrated our workflow into an R package cosmiq, which can be well integrated with the popular software xcms. For a joined analysis of multiple LCMS runs, we demonstrate detection of up to ten times more features with a higher evectiveness compared to classical processing approaches. Therefore we recommend this approach especially for LCMS

based metabolomics applications where a large amount of samples have to be analysed.

3 Metabolic alterations during colorectal cancer development – a combined analysis of the metabolome and transcriptome

David Jonathan Fischer[1], Christine Manser[3], Peter Bauerfeind[3], Giancarlo Marra[2] and Endre Laczko[1]

[1]Functional Genomics Center Zurich and [2]Institute of Molecular Cancer Research, University of Zurich, 8057 Zurich, Switzerland and
[3]Division of Gastroenterology and Hepatology, University Hospital Zurich, Zurich, Switzerland

D.J. Fischer contributed to the design of the experiments, analysed the data, and wrote the manuscript. C. Manser and P. Bauerfeind were involved in collection of biopsy samples. E.Laczko and G. Marra supervised the study and contributed to the experimental design.

3.1 Abstract

Reprogramming of metabolism has been considered as hallmark of many different cancer types. High throughput technologies such as metabolomics are therefore invaluable to identify novel metabolic pathways or characterize existing pathways in established cancer models where metabolism is less investigated. Our current knowledge on cancer metabolism mainly originates from studies on cell lines or cancer biopsies whereas premalignant lesions are less understood in terms of metabolic reprogramming. In this study we provide an insight into metabolic alterations of neoplastic colorectal lesions and cancer by a combined analysis of the metabolome, lipidome and transcriptome of human biopsies. We show that enzymes and metabolites involved in glycolysis, nucleotide, serine, glutamine and proline metabolism are similarly deregulated in adenoma and cancer, suggesting that reprogramming of these pathways is an early event in colorectal tumorigenesis. These early metabolic changes might be mediated by the transcription factor *MYC*, because the expression of *MYC*-controlled genes and genes involved in these metabolic pathways are closely correlated. We further identify cancer specific expression of solute carrier genes (SLC) which are important for the intake of glucose and neutral amino acids, while transporter for cationic amino acids are upregulated in both lesions. The exclusive overexpression of some transporter genes in cancer and not in adenoma might explain our observation that some metabolites are only increased in cancer. Lipidome analysis revealed that many lipids with very long chained, polyunsaturated fatty acids are increased while low chained, saturated fatty acids are decreased in adenomas and cancer. The upregulation of genes involved in fatty acid synthesis, elongation and desaturation in these tissue support a scenario where these lipid classes are synthesized de novo.

3.2 Introduction

The study of metabolism in cancer research has gained much attention during the last years. Recently, the altered metabolic phenotype of cancer cells was added to the list of prominent hallmarks of cancer (Hanahan and Weinberg, 2011). One of the well-known discoveries in this field is an increased aerobic production of lactic acid in cancerous versus normal tissue and was first described anno 1926 by Otto Warburg (Warburg, 1926). During the last

decade of research it became evident that cancer metabolism is linked to genetic alterations of molecular signalling pathways. For example, alterations in core metabolism were shown to be associated with activated oncogenes and mutant tumor suppressor genes (Vander Heiden et al., 2009). The transcription factor MYC has been shown to be responsible for the reprogramming of a broad range of different metabolic reactions (Li and Simon, 2013). Nowadays, a diverse spectrum of metabolic pathways beyond aerobic glycolysis are known to be characteristic for cancer cells. Among these pathways are several biosynthetic reactions of lipids, nucleotides, serine or proline (Benjamin et al., 2012; Liu et al., 2012). In addition to an increased demand for glucose, cancer cells also rely on glutamine as an important fuel for proliferation (Wise et al., 2008). In many cases where altered cancer metabolism was reported, the metabolic effect was identified as a growth advantage in order to allow cancer cells a rapid proliferation (Nomura et al., 2010; Possemato et al., 2011; Son et al., 2013). However, inactivation of the APC gene and subsequent overexpression of the *MYC* gene was described as a key event of colorectal tumorigenesis and therefore *MYC* is already upregulated in pre-cancerous lesions (Sansom et al., 2007). Based on this premise, we hypothesized whether altered metabolism might not be only cancer-specific but already characteristic for early non-cancerous neoplastic lesions. In this study we therefore aimed to characterize metabolic alterations in benign colorectal adenomas and cancer by investigating the metabolome (metabolites and lipids) and transcriptome of human colorectal tissue. The metabolome is the ultimate biochemical output of the genome, transcriptome and proteome (Patti et al., 2012). A combined analysis of metabolomics data with other omics data can therefore improve our understanding of biological processes, such as metabolic adaptions to environmental changes (Kresnowati et al., 2006). Application of clustering and multivariate methods to transcriptomics and metabolomics data was employed to identify condition dependent links between specific genes and metabolites in microorganisms (Jozefczuk et al., 2010). These strategies might be also successful to better understand human malignancies such as colorectal cancer, and especially cancer metabolism. Metabolomics alone was already successfully applied to colorectal biopsies in order to identify metabolite alterations between cancer and normal mucosa (Denkert et al., 2008; Chan et al., 2009; Hirayama et al.,

2009; Ong et al., 2010). These studies already highlighted deregulated metabolites involved in glycolysis, pentose phosphate pathway and amino acid and nucleotide metabolisms. With the hereby-presented study we want to provide a better insight into early metabolome changes in colorectal cancer development, notably in the pre-cancerous adenomas. In addition, we aim to identify the enzymatic reactions that are behind the metabolic reprogramming by integrating metabolomics and transcriptomics data.

3.3 Materials and Methods

3.3.1 Ethics approval

The study was approved by the local ethics commission of the Kanton Zürich and written informed consent was obtained from each participating patient.

3.3.2 Collection of biopsies

Tissue biopsies of colorectal neoplasms and adjacent normal mucosa were collected during routine colonoscopy procedures. In order to avoid sampling of non-tumourous tissue, only a small sample (3-5 mg) from the tip of the tumor was taken. Only polyps with at least 10 mm diameter were considered to ensure that the sampling procedure would not interfere with the histological examinations. Lesions of the classes polypoid and non-polypoid were included to the study, according to the Paris classification system (Participants in the Paris Workshop, 2003). For the collection of normal mucosa, a small biopsy was taken from the healthy mucosa of the same colon segment with at least 2 cm space to the site of the lesion. The biopsy was immediately frozen in liquid nitrogen after removal and stored at -80° C until further usage.

3.3.3 Metabolome and lipidome extraction and analysis

The fresh weight of each frozen biopsy was determined and extraction solvent was added at a ratio of 100 µl per 3 mg biopsy. The extraction solvent consisted of 98% methanol, 1% metabolite internal standard solution and 1% lipid internal standard solution (See table A-1). Immediately after solvent addition the tissue was mechanically lysed in a glass homogenizer (Dounce homogenizer) on ice. The homogenate was centrifuged for 5 minutes at 4000 g and the clear methanol extract was removed for further analysis. For the

lipidomics and metabolomics analysis, we employed two capLCMS methods using self-packed capillary columns. For the metabolome analysis we used a HILIC method at basic conditions (pH9) and for the lipidome analyses we used a reversed phase (RP) method. For a detailed description of the methods we refer to table 1.1 in chapter 1.4.3. For the lipidome analyses we diluted the methanol extracts five times with buffer A of the RP method and transferred them to the autosampler vials for LCMS analysis. For metabolome analyses we dried the methanol extract using a vacuum concentrator (speedvac) and re-dissolved the dried metabolites in ultrapure water. This aqueous solution was diluted five times with the injection solvent mixture consisting of 50 mM ammonium acetate, 90% acetonitrile, 9% methanol and 1% ultrapure water adjusted to pH 9 with concentrated ammonia solution. Before LCMS analysis, the processed samples were centrifuged for 2 minutes at 4000 g to remove residual precipitates. MS analysis was performed on a waters Synapt G2 HDMS mass spectrometer using positive (RP) or negative (HILIC) mode electrospray ionization (ESI+ or ESI-) at a scan rate of 0.3 scans per seconds (3Hz) in MSE mode (all ion fragmentation mode). Detailed MS tune settings were as described in section 1.3.3.

3.3.4 LCMS data processing

Continuum LCMS data were first centroided and then converted to vendor independent netCDF format using Waters MassLynx 4.1 and the integrated DataBridge software. The netCDF centroid data were then processed using the R package cosmiq, which was described in chapter 2. The following tuning parameters for the processing with cosmiq were used: Signal-to-noise ratio (SNR) of mass peak detection was set to 3, SNR for chromatographic detection to 20 and m/z bin size was set to 0.003 Da. The resulting peak table was first normalized by means of dividing the intensity values of each sample by the median intensity of selected internal standards (See appendix table A-1). As a second step, each intensity value was log(2) transformed. The normalized data table was then matched to a library of metabolite standards with known retention time by mass and retention using a retention time cutoff of 20 seconds and a mass cutoff of 0.01 Da. For a list of these standards we refer to chapter 1, Table 1.3. For the annotation of unknown metabolites we compared the accurate mass against either the KEGG (Ogata et al., 1999)

or the HMDB database (Wishart et al., 2013). For the HILIC dataset we searched for possible hits assuming a [M-H]- ion. For the RP dataset we assumed [M+H]+ or [M+NH4]+ as possible adducts. In both cases we annotated database hits within a mass accuracy tolerance of 0.01 Da.

3.3.5 Transcriptome data

Transcriptomic data based on Affymetrix Exon 1.0 microarrays from our past studies were included in this work (Cattaneo et al., 2011; Maglietta et al., 2012). Both studies contain data from colorectal biopsies of different tumor stages. We included the datasets of polypoid adenomas (n = 17), cancers (n = 17), and adjacent normal mucosa (n = 34). Clinical and histological details of these cases are reported in the two references mentioned above. Both datasets are publicly available in the ArrayExpress database under the accession numbers E-GEOD-21962 and E-MTAB-829 (www.ebi.ac.uk/arrayexpress). The raw chip data were parsed using the most recent probeset information (Entrez gene ID, version 18) of the brainarray database (Dai et al., 2005), yielding a total number of 24681 probesets. Log(2) expression values were obtained using the robust multiple-chip analysis (RMA) algorithm (Irizarry, 2003) and data was quantile normalized.

3.3.6 Data analysis: statistics

Large experimental datasets typically contain much more variables than observations and multivariate methods such as principal components analysis (PCA) are therefore not ideal to treat such data. Culhane et al. (2002) developed a classification method, which builds on ordination techniques such as PCA but circumvents the problem of having excess variables by ordinating the data based on predefined groups. This so called between group analysis (BGA) projects the data on N – 1 eigenvectors, where N is the total number of defined groups and each eigenvector distinguishes between two of these groups. The BGA function is available as part of the R package made4 (Culhane et al., 2005). Made4 was developed for the multivariate analysis of microarray data and mainly builds on the data ordination functions of another R package, ade4 (Thioulouse et al., 1997). We performed BGA in order to investigate the reflection of clinical phenotypes such as tumor stage or tumor size on the level of the three different omics datasets

(metabolomics, lipidomics, transcriptomics). Additionally we used BGA to classify genes, lipids and metabolites according to their expression level within these clinical phenotypes. The quantitative information for each variable (i.e. genes, metabolites and lipids) is transformed into the ordinated space such that a narrow angle between an ordinated variable and a certain group implies a good correlation between variable and group. Our classification approach therefore was to assign each variable to a group by selecting the closest angle between all group center coordinates and variable coordinates. The length of the ordination vector implies the contribution to each group, i.e. how significant each variable is altered within the group. In order to filter only for significant genes, metabolites or lipids, we applied a simple permutation test: The BGA was repeated 100 times with either of the three omics data matrices where each time the group label was randomly permuted. From the obtained random variable ordinations we calculated the euclidean distance between origin and the position at the first two discriminating axes. From this vector length distribution we chose the 99% percentile as a threshold for the ordinations in the non-permuted dataset. Each gene, lipid or metabolite with a length above this 99% percentile were considered as significant (See Appendix Figure A-6).

3.3.7 Data analysis: Biological interpretation

For the biological interpretation of the data we chose different approaches. First we performed a combined pathway over representation analysis of the metabolome and transcriptome data in order to investigate impacts in metabolic pathways. This analysis was done by parsing a list of annotated metabolites and genes to the webserver tool IMPaLA (Kamburov et al., 2011). The resulting list of pathways including enrichment analysis scores was exported and filtered for KEGG metabolic pathways only. For a more detailed investigation of single pathways we created networks based on KEGG pathway maps (Ogata et al., 1999). We obtained KGML files for each pathway from the KEGG homepage (www.genome.jp/kegg/pathway.html) and generated networks using cytoscape (Shannon et al., 2003). Each metabolite was represented as node and for each enzymatic reaction with one or more annotated human genes an edge was created between the nodes. Transcriptomics data were mapped on each edge and metabolomics data on

each node using color codes which represent the grouping from the BGA analysis. For some reactions there are multiple genes annotated, for example the five isoforms of the enzyme lactate dehydrogenase (LDHC, LDHB, LDHA, LDHAL6A, LDHAL6B). Metabolic pathways in the KEGG annotation however cannot distinguish between such isoforms. Our custom metabolic network maps have the advantage that each gene is individually displayed, which is important for cases where gene isoforms show a different expression pattern. For the analysis of solute carrier transporters (SLC), we annotated metabolite substrates based on the bioparadigms database, a web resource for existing knowledge about SLC genes and classes (www.bioparadigms.org) (Hediger et al., 2013).

3.4 Results and Discussion

3.4.1 Group classification reveals alterations in the metabolome, lipidome and transcriptome of different colorectal cancer progression stages

The raw data processing yielded a total number of 24681 annotated gene transcripts, 23912 and 45714 non-annotated mass/retention time features for the metabolomics or lipidomics dataset, respectively. The metabolomics and lipidomics data in this work was obtained from a total of 49 paired colorectal lesions and normal mucosa biopsies (total, 98 tissue samples) from 39 patients (Table 3.1). The aim of this study was to interpret metabolic changes by integrating the molecular changes observed in all three omics datasets. Since the transcriptomics study was performed with a completely different patient cohort than in this study, we aimed to integrate the data according to clinical paramaters. Among the clinical factors that were annotated in both studies are lesion diameter, patient age, tissue type, degree of dysplasia, microscopic appearance or Paris classification (See Table 3.1 and tables of references (Cattaneo et al., 2011; Maglietta et al., 2012)). A BGA was calculated on the metabolomics dataset using different clinical parameters as grouping factors (Figure A-1). Based on this analysis, the grouping factor “tissue type” revealed the highest number of statistically altered metabolites between the groups (Figure A-1 c). Other factors, such as lesion diameter, patient age, Paris classification or degree of dysplasia (Appendix Figure A-1 b,f,d respectively) showed separation of the group clusters but only a low number of significant metabolite features, according to a 99% confidence threshold

from randomly permuted data (See materials and methods). The finding that tissue type was the most dominant discriminating factor to explain variation in the metabolome supported our observations from previous studies on the transcriptome, where the highest degree of gene expression variation was explained between normal mucosa, polypoid or non-polypoid lesions, or cancer (Sabates Bellver et al., 2007; Cattaneo et al., 2011). Motivated by this result we decided to apply BGA analysis on all omics data using the tissue types normal mucosa (NM), polyps (P) and cancer (CA) as grouping factors. The result of the BGA is shown in Figure 3.1 for metabolomics (Fig. 3.1 a), lipidomics (Fig 3.1 b) and transcriptomics (Fig 3.1 c) data. In every case, the first ordination axis separated between the groups NM and the two lesion types. For the metabolomics data the amount of explained variance was 77.7% for the first axis, indicating that the largest amount of metabolome changes account for the difference between normal mucosa and lesion. The second axis separated the two lesion types P and CA. For the transcriptomics dataset there was a clear separation of the three groups, whereas in the metabolomics and lipidomics dataset the NM and P group are overlapping. Interestingly, the CA cluster is further apart from the NM cluster on axis 1 in all datasets, indicating that common alterations in both lesion types are more discinct in the cancer tissue. For the analysis of metabolomics and other high throughput data it is important to validate classification methods, especially if supervised classification methods like BGA are used (Kjeldahl and Bro, 2010). In order to identify significant alterations in each of the omics data, we applied a permutation test (See material and methods). As the length of the vector in the n-dimensional ordination space implies its contribution to a certain group (Culhane et al., 2002), the coordinates of the first two axes were transformed to euclidean coordinates. All variables with a vector length above a 99% confidence threshold from the permutation test were considered as significant. This threshold is illustrated as red circle in Figure 3.1 in the right panel. As expected from the better separation of the groups NM, P and CA in the transcriptomics data, the amount of significantly differentially expressed genes is higher (36.6%) compared to altered metabolites (23.7%) or lipids (11.3%). To further investigate these lipid, metabolite and gene expression alterations, we classified each variable based on the three groups NM, P and CA (Figure 3.2). In the transformed ordination space, the variables which are

closely correlated to a certain group, lie in the same direction (Culhane et al., 2002). Conversely, negatively correlated variables and groups lie in the opposite direction. We therefore assigned each variable to the group center with the closest angle (See Figure 3.2 left panel). In total we assigned all transcripts, metabolites and lipids according to six classes: Up- or downregulated in both CA and P (PCA up, PCA down, blue color), up- or downregulated in CA (CA up, CA down, red color) and up- or downregulated in P (P up, P down, yellow color). The resulting gene expression clusters are illustrated on the right panel of Figure 3.2 as relative mean intensity values for each class. Figure A-2 illustrates example genes from each classification cluster of the transcriptomics data. The class with the highest amount of the genes, metabolites and lipids was PCA up or PCA down (45, 55 and 51.7% respectively), indicating that a major part of the metabolic alterations is common to cancer and adenoma.

Table 3.1: Clinical parameters of patients involved in the metabolomics and lipidomics study. Lesions from a total of 39 patients were collected. na = no data available.

Patient	Age	Sex	Colon segment involved	Maximum lesion diameter (mm)	Paris classification #	Microscopic appearance	Highest degree of dysplasia in the lesion ▪	No. of lesions at study colonoscopy	No. of previously excised lesions ‡
1	65	M	C	10	Iia-IIb	TVA	HGD	3	0
2	na	na	na	na	na	na	na	na	na
3	61	M	A	10	Is	TA	LGD	3	0
4a	57	M	A	40	Is	TA	LGD	5	2
4b			D	30	Ip	TVA	CA		
4c			S	40	Ip	TVA	CA		
5	49	M	S	30	Is	TA	LGD	7	0
6a	55	M	A	12	Ip	TA	LGD	3	0
6b			D	10	Ip	SA	na		
6c			S	25	Ip	TVA	HGD		
7	67	M	S	10	IIa	TVA	HGD	4	0
8	76	F	C	18	Iib-IIc	TVA	HGD	1	0
9	53	M	LF	30	Ip	TA	HGD	1	0
10	61	M	C	25	Is	TVA	LGD	0	0
11	54	M	D	10	Ip	na	na	2	1
12	43	M	S	26	Ip	TA	HGD	1	0
13	71	M	HF	40	Is	CA	na	6	0
14	67	M	T	12	Is	TVA	HGD	2	0
15	57	M	S	23	Ip	TVA	Tis	4	1
16a	69	M	S	20	Ip	TVA	HGD	6	0
16b			S	20	Ip	TVA	HGD		
17	57	M	S	20	Ip	TA	HGD	1	0
18a	74	M	HF	15	IIa	TA	HGD	na	na
18b			S	40	Is	TVA	HGD		
18c			S	30	Is	TVA	HGD		
19	56	M	S	27	Is	TA	HGD	1	0
20	na	na	na	na	na	na	na	na	na
21	53	M	S	10	Is	TA	HGD	3	0
22	55	M	C	40	Iia-IIb	TVA	LGD	1	0
23a	73	M	C	25	Is	TA	LGD	4	0
23b			A	30	Ip	TA	LGD		
24	63	M	T	8	IIa	TA	LGD	4	0
25	75	M	R	12	Is	TVA	LGD	0	6
26	na	na	na	na	na	na	na	na	na
27	49	M	S	15	IIa	na	na	1	0
28	72	M	S	20	Ip	TVA	LGD	2	0
29	44	F	S	10	Is	na	na	3	0
30	73	M	HF	20	Is	TVA	HGD	4	3
31	74	M	D	15	Is	TVA	LGD	6	3
32	54	F	R	25	Is	SA	HGD	3	0
33	48	F	S	40	Is	TA	CA	1	0
34	na	na	na	na	na	na	na	na	na
35a	77	F	C	30	IIa	TVA	HGD	5	0
35b			C	30	IIa	TVA	HGD		
35c			A	30	IIa	na	na		
36	na	na	na	na	na	na	na	na	na
37	75	F	na	40	na	CA	CA	na	na
38	62	M	na	40	na	CA	CA	na	na
39	58	M	S	25	Ip	TA	HGD	0	0

Abbreviations: M, male; F, female; C, caecum; A, ascending colon; HF, hepatic flexure; T, transversum; D, descending colon; S, sigma; R, rectum; TA, tubular adenoma; TVA, tubulovillous adenoma; VA, villous adenoma; MVSP, microvescicular serrated polyp; SA, serrated adenoma; SSA, sessile serrated adenoma; LGD, low-grade dysplasia; HGD, high-grade dysplasia.

Macroscopic appearance of neoplastic lesions was classified according to Paris Endoscopic Classification. The Paris Endoscopic Classification of Superficial Neoplastic Lesions. Gastrointest Endosc 2003;58(suppl.):S3-S27

▪ Low-grade versus high-grade dysplasia as defined by the WHO classification of tumors of the digestive system, editorial and consensus conference in Lyon, France, November 6-9, 1999.IARC

‡ Total no. adenomas detected and excised during previous colonoscopies.

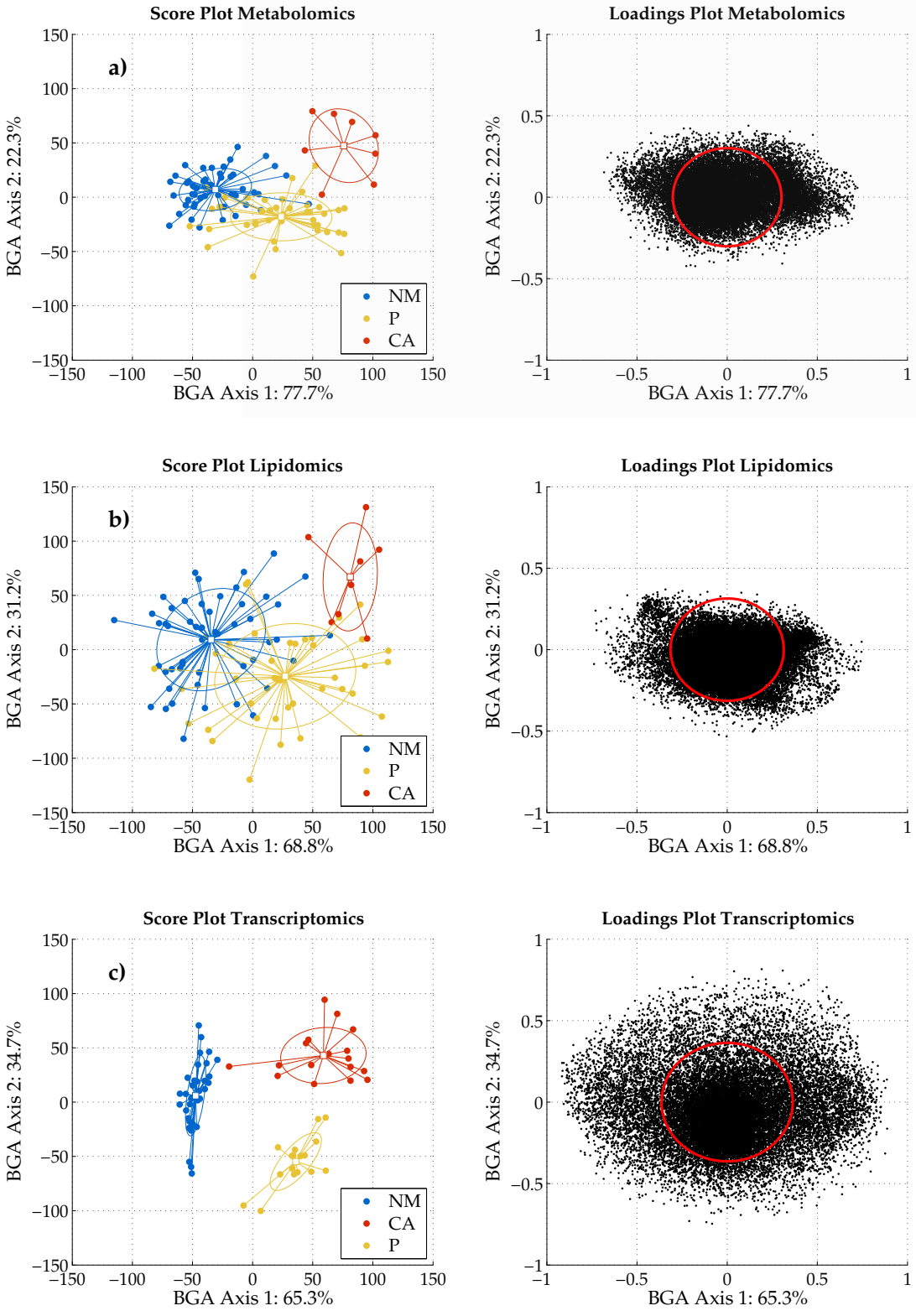


Figure 3.1: Between group analysis (BGA) of metabolomics (a), lipidomics (b) and transcriptomics data (c). The BGA was calculated using the group factor "tissue type" with the levels: normal mucosa (NM), polyp (P) and cancer (CA). The left panel shows the score plot (samples) of the first two BGA axes. The group center for each group is indicated by a square and a 50% confidence ellipse is shown for each group. The right panels shows the loadings plot (variables). The red ellipse in the loadings plot illustrates the 99% percentile of randomly generated loadings (See material and methods for details). The amount of variance for each BGA axis is given as percent of total variance.

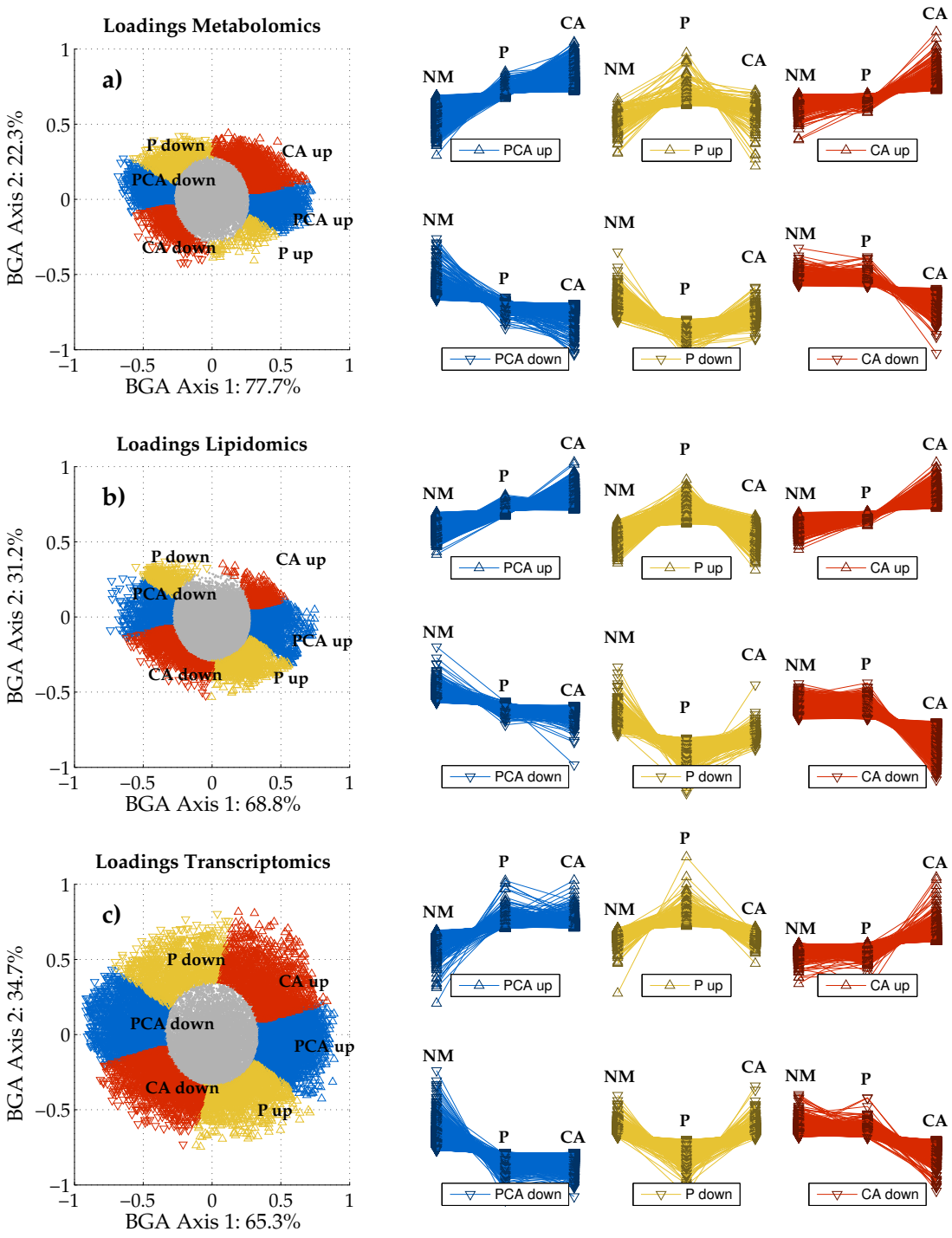
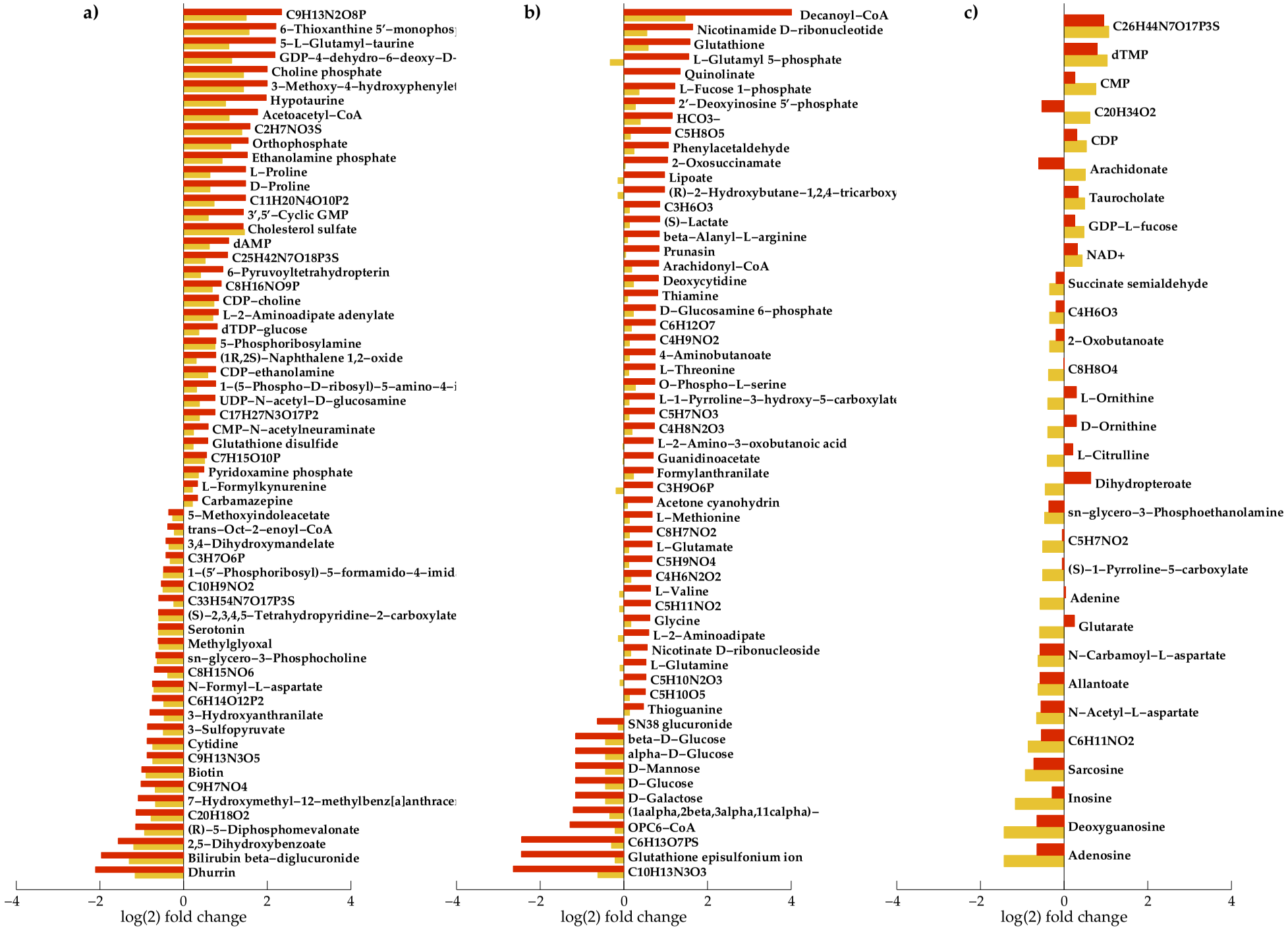


Figure 3.2: Classification of the metabolome (a), lipidome (b) and transcriptome (c) according to the between group analysis (BGA) and the grouping information for normal mucosa (NM), polyp (P) and cancer (CA). The left panel shows the loadings plot as in Figure 3.1, but highlights the loadings according to the class they were assigned to: Blue: deregulated in polyps and cancer (PCA up/PCA down), yellow: Deregulated only in polyps (P up/ P down) and red: deregulated in cancer only (CA up/ CA down). Grey: Loadings inside the confidence ellipse. Up/down indicates whether transcripts, metabolites or lipids are increased or decreased, respectively. The right panel shows relative mean intensity values for each metabolite with significant contribution.

3.4.2 Alteration of metabolic pathways

A total of 187 annotated metabolites were significantly changed and classified according to the BGA analysis. For each of these metabolites we calculated log(2)-fold changes between their abundance in the lesion and normal mucosa (Figure 3.3). 77 were classified as altered in PCA (Fig. 3.3 a), 75 as altered in CA (Fig. 3.3 b) and 35 as altered in P (Fig 3.3 c). While half of the metabolites in the class PCA were increased in cancer and polyps and the other half decreased, most of the metabolites in the class CA were increased. Interestingly, for almost all metabolites in the class PCA the fold change was higher in cancer compared to adenomas. Also, most metabolites in the class CA show a similar alteration trend for the adenomas as in cancer. For example, decanoyl-CoA with a high fold change of 4.1 in cancer is also increased by a lower fold change of 1.4 in adenomas (Fig 3.3 b). The fold changes in the P class are generally lower compared to the classes PCA and CA. Also, the fold change difference between adenoma and cancer is in many cases quite low. Interestingly, some compounds in this class show completely different trends in fold change. For example, arachidonate is increased in adenomas, wheras it is decreased in cancer. Overall, most metabolite fold changes show a similar trend in adenomas and cancer, whereas the amount of fold change is considerably greater in cancer. Taken together, these results implicate a nearly overlapping alteration in metabolism for both adenoma and cancer. However it seems that the amount of variation is more advanced in cancer.

Figure 3.3 (next page): Metabolite fold changes comparing cancer (CA) versus normal mucosa (NM) in red, or adenoma (P) versus normal mucosa (yellow). The metabolites were classified according to BGA analysis (See Figure 3.2): a) deregulated in adenoma and cancer, b) deregulated in cancer and c) deregulated in adenoma.


a)
C9H13N2O8P
6−Thioxanthine 5′−monophos
5−L−Glutamyl−taurine
GDP−4−dehydro−6−deoxy−D-
Choline phosphate
3−Methoxy−4−hydroxyphenylet
Hypotaurine
Acetoacetyl−CoA
C2H7NO3S
Orthophosphate
Ethanolamine phosphate
L−Proline
D−Proline
C11H20N4O10P2
3′,5′−Cyclic GMP
Cholesterol sulfate
dAMP
C25H42N7O18P3S
6−Pyruvoyltetrahydropterin
C8H16NO9P
CDP−choline
L−2−Aminoadipate adenylate
dTDP−glucose
5−Phosphoribosylamine
(1R,2S)−Naphthalene 1,2−oxide
CDP−ethanolamine
1−(5−Phospho−D−ribosyl)−5−amino−4−i
UDP−N−acetyl−D−glucosamine
C17H27N3O17P2
CMP−N−acetylneuraminate
Glutathione disulfide
C7H15O10P
Pyridoxamine phosphate
L−Formylkynurenine
Carbamazepine
5−Methoxyindoleacetate
trans−Oct−2−enoyl−CoA
3,4−Dihydroxymandelate
C3H7O6P
1−(5′−Phosphoribosyl)−5−formamido−4−imid
C10H9NO2
C33H54N7O17P3S
(S)−2,3,4,5−Tetrahydropyridine−2−carboxylate
Serotonin
Methylglyoxal
sn−glycero−3−Phosphocholine
C8H15NO6
N−Formyl−L−aspartate
C6H14O12P2
3−Hydroxyanthranilate
3−Sulfopyruvate
Cytidine
C9H13N3O5
Biotin
C9H7NO4
7−Hydroxymethyl−12−methylbenz[a]anthrace
C20H18O2
(R)−5−Diphosphomevalonate
2,5−Dihydroxybenzoate
Bilirubin beta−diglucuronide
Dhurrin
−4
−2
0
2
4
log(2) fold change
b)
Decanoyl−CoA
Nicotinamide D−ribonucleotide
Glutathione
L−Glutamyl 5−phosphate
Quinolinate
L−Fucose 1−phosphate
2′−Deoxyinosine 5′−phosphate
HCO3−
C5H8O5
Phenylacetaldehyde
2−Oxosuccinamate
Lipoate
(R)−2−Hydroxybutane−1,2,4−tricarboxy
C3H6O3
(S)−Lactate
beta−Alanyl−L−arginine
Prunasin
Arachidonyl−CoA
Deoxycytidine
Thiamine
D−Glucosamine 6−phosphate
C6H12O7
C4H9NO2
4−Aminobutanoate
L−Threonine
O−Phospho−L−serine
L−1−Pyrroline−3−hydroxy−5−carboxylate
C5H7NO3
C4H8N2O3
L−2−Amino−3−oxobutanoic acid
Guanidinoacetate
Formylanthranilate
C3H9O6P
Acetone cyanohydrin
L−Methionine
C8H7NO2
L−Glutamate
C5H9NO4
C4H6N2O2
L−Valine
C5H11NO2
Glycine
L−2−Aminoadipate
Nicotinate D−ribonucleoside
L−Glutamine
C5H10N2O3
C5H10O5
Thioguanine
SN38 glucuronide
beta−D−Glucose
alpha−D−Glucose
D−Mannose
D−Glucose
D−Galactose
(1aalpha,2beta,3alpha,11calpha)−
OPC6−CoA
C6H13O7PS
Glutathione episulfonium ion
C10H13N3O3
−4
−2
0
2
4
log(2) fold change
c)
C26H44N7O17P3S
dTMP
CMP
C20H34O2
CDP
Arachidonate
Taurocholate
GDP−L−fucose
NAD+
Succinate semialdehyde
C4H6O3
2−Oxobutanoate
C8H8O4
L−Ornithine
D−Ornithine
L−Citrulline
Dihydropteroate
sn−glycero−3−Phosphoethanolamine
C5H7NO2
(S)−1−Pyrroline−5−carboxylate
Adenine
Glutarate
N−Carbamoyl−L−aspartate
Allantoate
N−Acetyl−L−aspartate
C6H11NO2
Sarcosine
Inosine
Deoxyguanosine
Adenosine
−4
−2
0
2
4
log(2) fold change

In order to interpret the metabolic changes in adenomas and cancer, we next performed a pathway overrepresentation analysis. Both transcriptomics and annotated metabolomics data were parsed to the web-based IMPaLA tool (Kamburov et al., 2011). For each of the BGA classes we retrieved a list of significantly altered metabolic pathways (Table 3.2). Interestingly, the analysis revealed some metabolic pathways to be common for polyps and cancer which are known to be involved in metabolic reprogramming of cancer (Nucleotide metabolism, Pentose phosphate metabolism, glycolysis, fatty acid biosynthesis) (DeBerardinis et al., 2008). In cancer, these pathways are together involved in an increased uptake of glucose in order to provide enough energy and carbon backbone structures for the production of biomass (DNA, RNA and lipids). The overrepresentation of these pathways in both cancer and polyps points to a metabolic alteration which is already present in early lesions and persist when the polyps progress to cancer. Other pathways such as arginine/proline metabolism or glycine/serine threonine seem to be differently altered in the polyps or in cancer. The results of the pathway analysis are together highlighted on a global metabolic pathway map (Figure 3.4 for the transcriptomics data and Figure 3.5 for metabolomics data). Interestingly, in most of the pathways the predominating genes are in the PCA class whereas for the metabolites most are in the CA class. This observation indicates that most alterations in enzyme gene expression level may be shared by adenomas and cancer whereas many metabolites are only altered in cancer. The map in Figure 3.4 indicates that many alterations in enzyme gene expression are spread across many different kinds of metabolic pathways, whereas the metabolite changes are more dominating around central carbon, nucleotide and amino acid metabolism (Figure 3.5).

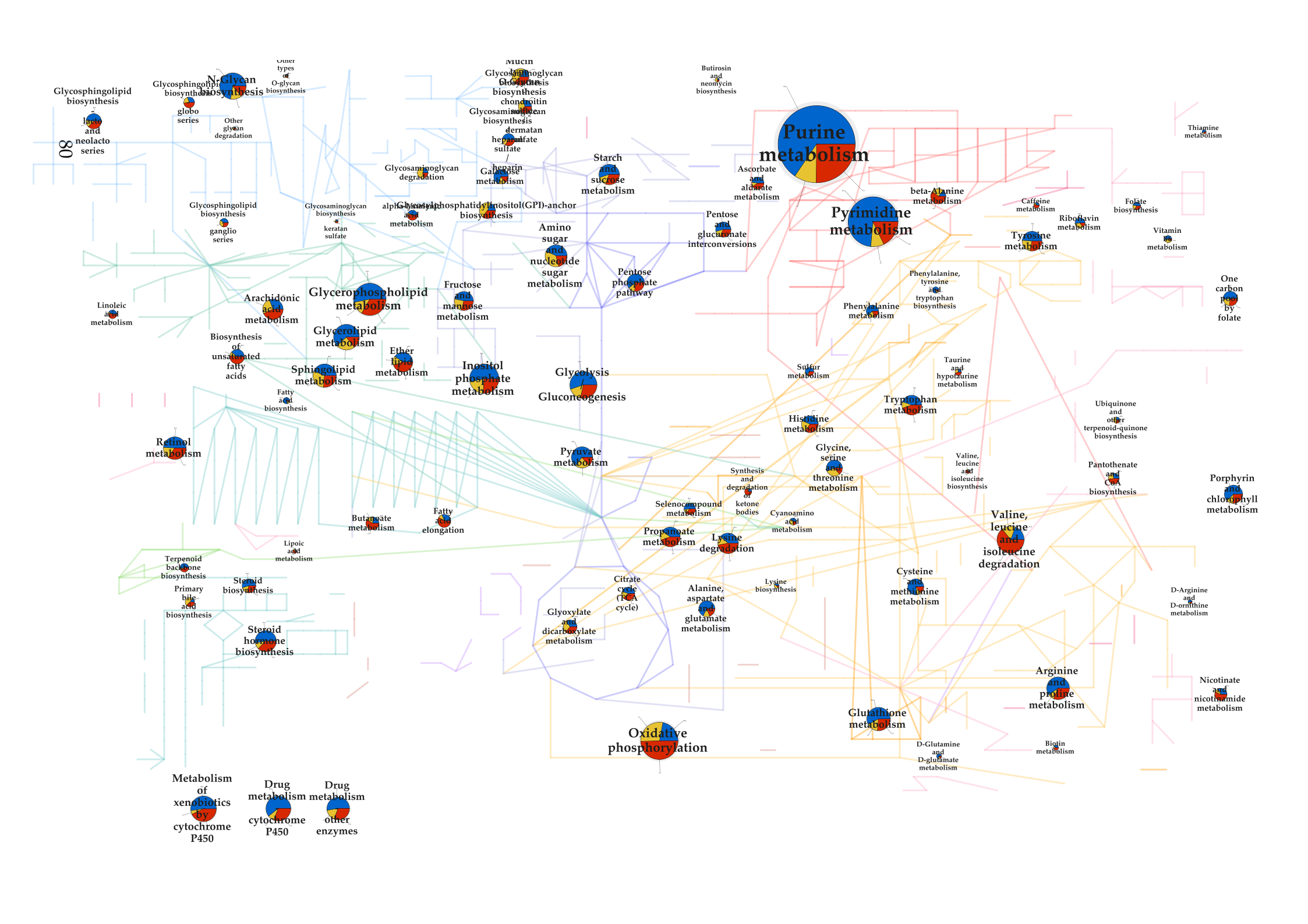
Table 3.2: Integrated pathway overrepresentation analysis of metabolomics and transcriptomics data using IMPaLA (Kamburov et al., 2011). Only KEGG pathways involved in metabolism and with *p-joint* value <= 0.05 were retrieved. P-values represent pathway significance according to hypergeometric distribution calculation (Cavill et al., 2011).

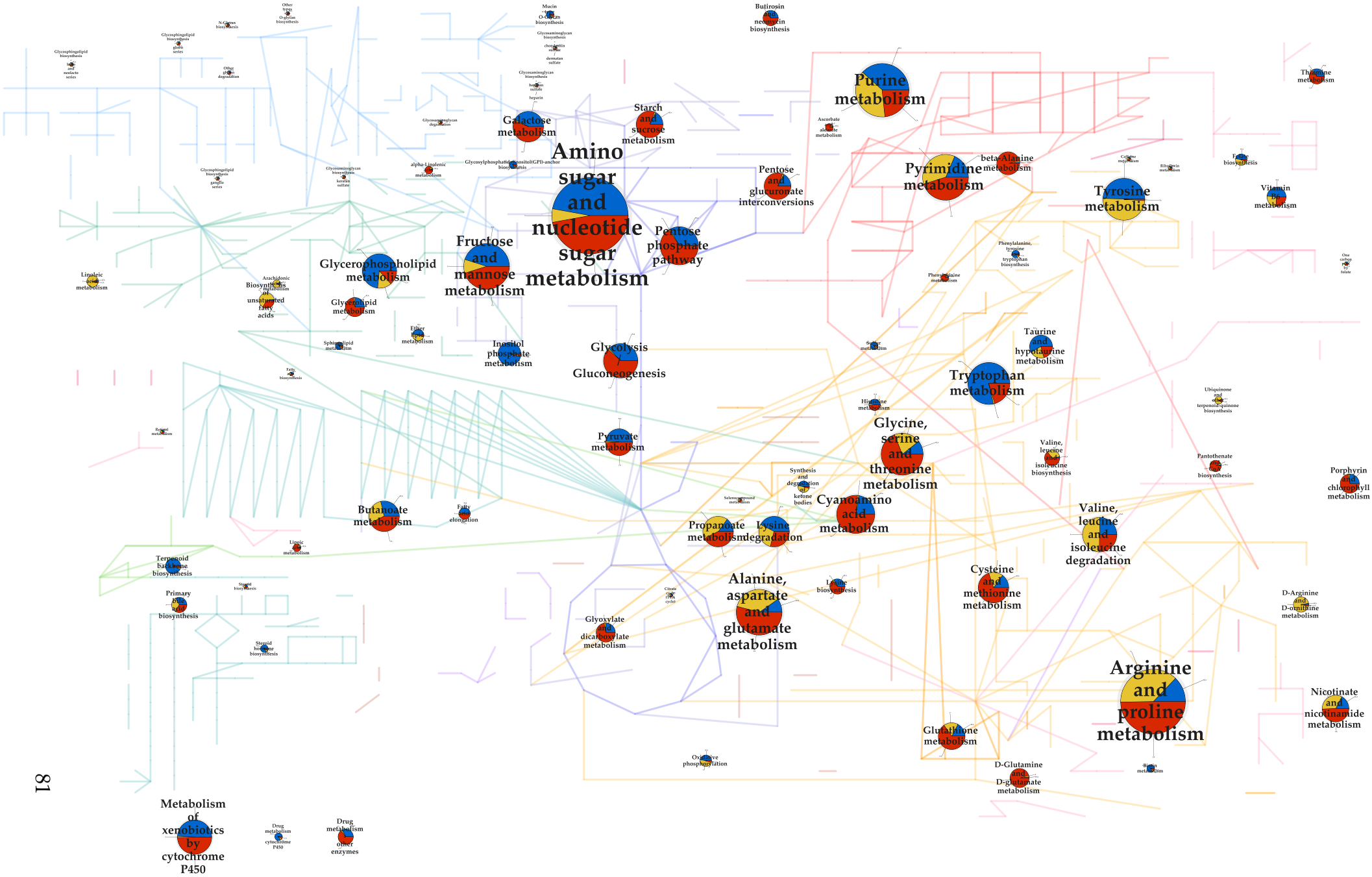
KEGG Pathway Name	# genes	*p* genes	# metabolites	*p* metabolites	*p* joint
Deregulated in polyps and cancer					
Purine metabolism	66	5.44E-07	5	0.0303	1.65E-08
Pyrimidine metabolism	47	1.54E-07	2	3.61E-01	5.57E-08
Amino sugar and nucleotide sugar metabolism	11	0.464	9	7.76E-05	3.60E-05
Tryptophan metabolism	10	0.381	8	1.40E-04	5.34E-05
Inositol phosphate metabolism	19	0.0615	5	1.36E-03	8.34E-05
N-Glycan biosynthesis	23	0.000093	0	1	0.000093
Glycerophospholipid metabolism	21	0.442	6	0.000405	0.000179
Pyruvate metabolism	16	0.0101	3	0.0224	0.000226
Pentose phosphate pathway	12	0.0108	3	0.0263	0.000284
Fructose and mannose metabolism	10	0.255	5	0.00265	0.000674
Galactose metabolism	11	0.0481	3	0.0426	0.00205
Tyrosine metabolism	11	0.223	5	0.0144	0.00321
Glycolysis / Gluconeogenesis	18	0.185	3	0.0206	0.0038
Phenylalanine. tyrosine and tryptophan biosynthesis	4	0.00962	1	0.473	0.00455
Taurine and hypotaurine metabolism	1	0.917	3	0.00797	0.0073
Glycerolipid metabolism	19	0.0221	1	0.473	0.0104
Cysteine and methionine metabolism	14	0.0161	1	0.67	0.0108
Alanine. aspartate and glutamate metabolism	12	0.0335	1	0.372	0.0125
Vitamin B6 metabolism	3	0.125	2	0.124	0.0155
Sulfur metabolism	5	0.0477	1	0.43	0.0205
Fatty acid biosynthesis	4	0.0238	0	1	0.0238
Phenylalanine metabolism	8	0.0281	0	1	0.0281
Terpenoid backbone biosynthesis	4	0.71	3	0.0453	0.0321
Glutathione metabolism	16	0.0771	1	0.522	0.0402
Lysine degradation	9	0.78	3	0.0598	0.0467
Deregulated in polyps					
Alanine. aspartate and glutamate metabolism	2	0.85	4	4.98E-05	4.23E-05
D-Arginine and D-ornithine metabolism	0	1	3	7.49E-05	7.49E-05
Arginine and proline metabolism	2	0.984	6	0.000117	0.000115
Tyrosine metabolism	5	0.363	5	0.000493	0.000179
Valine. leucine and isoleucine degradation	5	0.467	4	0.000426	0.000199
Purine metabolism	9	0.993	5	0.00119	0.00118
Propanoate metabolism	3	0.644	3	0.0038	0.00245
Glycosylphosphatidylinositol(GPI)-anchor biosynthesis	8	0.00246	0	1	0.00246
Pyrimidine metabolism	5	0.986	4	0.00259	0.00255
Glycosaminoglycan degradation	6	0.00918	0	1	0.00918
Mucin type O-Glycan biosynthesis	8	0.0104	0	1	0.0104
Linoleic acid metabolism	0	1	2	0.0253	0.0253
Glycine. serine and threonine metabolism	5	0.342	2	0.0752	0.0257
Amino sugar and nucleotide sugar metabolism	9	0.0443	1	0.585	0.0259
Vitamin B6 metabolism	2	0.117	1	0.25	0.0293
Primary bile acid biosynthesis	4	0.0862	1	0.345	0.0298
Butanoate metabolism	2	0.757	2	0.0511	0.0387

KEGG Pathway Name	# genes	*p* genes	# metabolites	*p* metabolites	*p* joint
Deregulated in Cancer					
beta-Alanine metabolism	10	3.65E-03	5	0.000258	9.40E-07
Alanine. aspartate and glutamate metabolism	3	0.831	6	4.09E-06	3.40E-06
Pentose phosphate pathway	5	0.332	6	0.0000351	1.17E-05
Cyanoamino acid metabolism	1	0.641	7	2.08E-05	1.33E-05
Glycine. serine and threonine metabolism	2	0.973	7	0.0000416	0.0000405
D-Glutamine and D-glutamate metabolism	0	1	4	0.0000544	0.0000544
Amino sugar and nucleotide sugar metabolism	5	0.786	9	0.0000776	0.000061
Arginine and proline metabolism	8	0.522	8	0.00027	0.000141
Glycolysis / Gluconeogenesis	9	0.554	5	0.000258	0.000143
Propanoate metabolism	10	0.00801	3	0.0305	0.000245
Butanoate metabolism	8	0.0188	3	0.0426	0.000801
Fructose and mannose metabolism	6	0.366	5	0.00265	0.000968
Pentose and glucuronate interconversions	5	0.502	5	0.00265	0.00133
Starch and sucrose metabolism	7	0.656	5	0.00265	0.00174
Biosynthesis of unsaturated fatty acids	8	0.00456	1	0.65	0.00297
Glutathione metabolism	7	0.554	4	0.00557	0.00308
Cysteine and methionine metabolism	2	0.966	5	0.00431	0.00417
Fatty acid elongation	8	0.00856	1	0.503	0.0043
Nicotinate and nicotinamide metabolism	7	0.0768	3	0.0598	0.00459
Galactose metabolism	2	0.93	4	0.00732	0.0068
Pyrimidine metabolism	11	0.864	5	0.00806	0.00696
Pantothenate and CoA biosynthesis	5	0.0706	2	0.0992	0.007
Lysine degradation	11	0.062	2	0.226	0.014
Valine. leucine and isoleucine biosynthesis	1	0.254	2	0.0705	0.0179
Tryptophan metabolism	10	0.0383	2	0.468	0.018
Pyruvate metabolism	3	0.932	3	0.0224	0.0209
Glyoxylate and dicarboxylate metabolism	5	0.221	3	0.0983	0.0218
Glycerolipid metabolism	4	0.953	3	0.0243	0.0232
Synthesis and degradation of ketone bodies	4	0.0245	0	1	0.0245
Glycosphingolipid biosynthesis - globo series	5	0.0323	0	1	0.0323
Lipoic acid metabolism	1	0.356	1	0.109	0.0389
Metabolism of xenobiotics by cytochrome P450	13	0.202	4	0.2	0.0403
Drug metabolism - other enzymes	8	0.285	2	0.17	0.0486

Figure 3.4 (page 80): Representation of metabolic pathways according to the number of altered transcriptome. Each circle represents a KEGG metabolic pathway. Circle size represents relative amount of deregulated genes within each pathway. The colored pie graphs indicate the amount of altered genes within one of the three BGA classes: blue: deregulated in adenoma and cancer, yellow: deregulated in adenoma, red: deregulated in cancer.

Figure 3.5 (page 81): Same as Figure 3.4 for the metabolomics data.


Purine metabolism
Pyrimidine metabolism
Glycerophospholipid metabolism
Oxidative phosphorylation
N-Glycan biosynthesis
Glycosphingolipid biosynthesis - lacto and neolacto series
Glycosphingolipid biosynthesis - globo series
Other types of O-glycan biosynthesis
Other glycan degradation
Mucin type O-glycan biosynthesis
Glycosaminoglycan biosynthesis - chondroitin sulfate / dermatan sulfate
Glycosaminoglycan biosynthesis - heparin sulfate / heparin
Glycosaminoglycan degradation
Galactose metabolism
Starch and sucrose metabolism
Butirosin and neomycin biosynthesis
Ascorbate and aldarate metabolism
beta-Alanine metabolism
Thiamine metabolism
Caffeine metabolism
Riboflavin metabolism
Folate biosynthesis
Vitamin B6 metabolism
Tyrosine metabolism
Glycosphingolipid biosynthesis - ganglio series
Glycosaminoglycan biosynthesis - keratan sulfate
alpha-Linolenic acid metabolism
Glycosylphosphatidylinositol(GPI)-anchor biosynthesis
Amino sugar and nucleotide sugar metabolism
Pentose and glucuronate interconversions
Pentose phosphate pathway
Phenylalanine, tyrosine and tryptophan biosynthesis
Phenylalanine metabolism
One carbon pool by folate
Linoleic acid metabolism
Arachidonic acid metabolism
Fructose and mannose metabolism
Biosynthesis of unsaturated fatty acids
Glycerolipid metabolism
Ether lipid metabolism
Sphingolipid metabolism
Inositol phosphate metabolism
Glycolysis / Gluconeogenesis
Fatty acid biosynthesis
Sulfur metabolism
Taurine and hypotaurine metabolism
Tryptophan metabolism
Histidine metabolism
Ubiquinone and other terpenoid-quinone biosynthesis
Retinol metabolism
Pyruvate metabolism
Glycine, serine and threonine metabolism
Valine, leucine and isoleucine biosynthesis
Pantothenate and CoA biosynthesis
Porphyrin and chlorophyll metabolism
Synthesis and degradation of ketone bodies
Selenocompound metabolism
Cyanoamino acid metabolism
Butanoate metabolism
Fatty acid elongation
Propanoate metabolism
Lysine degradation
Valine, leucine and isoleucine degradation
Lipoic acid metabolism
Terpenoid backbone biosynthesis
Steroid biosynthesis
Primary bile acid biosynthesis
Steroid hormone biosynthesis
Citrate cycle (TCA cycle)
Lysine biosynthesis
Cysteine and methionine metabolism
D-Arginine and D-ornithine metabolism
Alanine, aspartate and glutamate metabolism
Glyoxylate and dicarboxylate metabolism
Arginine and proline metabolism
Nicotinate and nicotinamide metabolism
Glutathione metabolism
D-Glutamine and D-glutamate metabolism
Biotin metabolism
Metabolism of xenobiotics by cytochrome P450
Drug metabolism - cytochrome P450
Drug metabolism - other enzymes


Amino sugar and nucleotide sugar metabolism
Purine metabolism
Pyrimidine metabolism
Tyrosine metabolism
Tryptophan metabolism
Glycine, serine and threonine metabolism
Cyanoamino acid metabolism
Alanine, aspartate and glutamate metabolism
Arginine and proline metabolism
Cysteine and methionine metabolism
Valine, leucine and isoleucine degradation
Valine, leucine and isoleucine biosynthesis
Lysine degradation
Propanoate metabolism
Butanoate metabolism
Glycolysis Gluconeogenesis
Pyruvate metabolism
Glyoxylate and dicarboxylate metabolism
Citrate cycle (TCA cycle)
Oxidative phosphorylation
Galactose metabolism
Starch and sucrose metabolism
Fructose and mannose metabolism
Pentose phosphate pathway
Pentose and glucuronate interconversions
Glycerophospholipid metabolism
Glycerolipid metabolism
Inositol phosphate metabolism
Ether lipid metabolism
Sphingolipid metabolism
Linoleic acid metabolism
Arachidonic acid metabolism
alpha-Linolenic acid metabolism
Fatty acid biosynthesis
Fatty acid elongation
Synthesis and degradation of ketone bodies
Taurine and hypotaurine metabolism
Phenylalanine metabolism
Phenylalanine, tyrosine and tryptophan biosynthesis
beta-Alanine metabolism
Ascorbate and aldarate metabolism
Pantothenate and CoA biosynthesis
Ubiquinone and other terpenoid-quinone biosynthesis
Porphyrin and chlorophyll metabolism
Nicotinate and nicotinamide metabolism
D-Arginine and D-ornithine metabolism
D-Glutamine and D-glutamate metabolism
Glutathione metabolism
Lysine biosynthesis
Thiamine metabolism
Riboflavin metabolism
Folate biosynthesis
Vitamin B6 metabolism
Caffeine metabolism
Sulfur metabolism
Selenocompound metabolism
Histidine metabolism
Butirosin and neomycin biosynthesis
Mucin type O-Glycan biosynthesis
Glycosaminoglycan biosynthesis - chondroitin sulfate / dermatan sulfate
Glycosaminoglycan biosynthesis - heparin sulfate / heparin
Glycosaminoglycan degradation
Glycosaminoglycan biosynthesis - keratan sulfate
Glycosylphosphatidylinositol(GPI)-anchor biosynthesis
Other types of O-glycan biosynthesis
N-Glycan biosynthesis
Glycosphingolipid biosynthesis - globo series
Glycosphingolipid biosynthesis - lacto and neolacto series
Glycosphingolipid biosynthesis - ganglio series
Other glycan degradation
Lipoic acid metabolism
Folate biosynthesis
One carbon pool by folate
Terpenoid backbone biosynthesis
Primary bile acid biosynthesis
Steroid biosynthesis
Steroid hormone biosynthesis
Biotin metabolism
Retinol metabolism
Lipoic acid metabolism
Metabolism of xenobiotics by cytochrome P450
Drug metabolism - cytochrome P450
Drug metabolism - other enzymes

3.4.3 Alteration of central metabolism is an early hallmark of colorectal cancer development

From the pathway overrepresentation analysis it became evident that central carbon metabolism, nucleotide metabolism and several amino acid metabolism pathways are altered based on differential expression of enzyme genes and metabolite changes. During the last years of research several metabolic alterations of cancer cells within these pathways were uncovered (DeBerardinis et al., 2008). Two main metabolic fuels, glucose and glutamine, support proliferating cells for energetic and biosynthesis processes. Beyond the well known Warburg effect, where glucose is converted to lactic acid, cancer cells utilize synthetic reactions for nucleotides, lipids, serine or proline (Benjamin et al., 2012; Li and Simon, 2013). We therefore investigated these metabolic pathways in more detail (Figure 3.6). Intriguingly, our gene expression data suggests that a major fraction of enzyme genes for these synthetic and energetically important reactions are upregulated in adenomas and cancer, i.e. all of these genes were classified as upregulated in PCA according to the BGA. We observe an increase in gene expression for *GPI*, *PFKM*, *GAPDH* or *PGK1*, indicating an increased glycolytic activity in adenomas and cancer. Other increased expressions for genes such as *G6PD*, *PGD*, *TKT* or *PRPS2* suggest that reactions of the pentose phosphate pathway are induced to generate 5-phosphoribosyl diphosphate (PRPP), an important precursor for nucleotide synthesis. An increased generation of nucleotides is also indicated by the expression of *UMPS*, *PPAT*, *CAD*, *DHODH* and *GART*. These genes are responsible for the first steps of purine and pyrimidine synthesis, involving *PRPP* and glutamine as metabolite precursors. In addition to the genes responsible for these initial reactions, most of the enzymatic genes responsible for interconversion of different purine und pyrimidine nucleotides are overexpressed as well (See Appendix Figure A-3 and A-4). Also important for the generation of purines is the incorporation of a glycine carbon during one-carbon metabolism reactions. Recent studies have shown that serine and subsequent glycine production is important for the rapid growth of cancer cells (Locasale et al., 2011; Possemato et al., 2011; Jain et al., 2012). Interestingly, these important genes for serine biosynthesis are overexpressed in our dataset for adenoma and cancer (*PSAT1*, *PSPH*, *SHMT2*). Another recent discovery in cancer metabolism is the synthesis of

proline from glutamine, which is believed to play an important role in redox balance (Liu et al., 2012; Phang et al., 2012). We found the relevant genes for these reactions to be overexpressed in our dataset (*GLS2*, *ALDH4A1*, *PYCR1*, *PYCRL*, *PYRC2*). Finally, another recent study report the KRAS mediated conversion of glutamine via aspartate and malate to pyruvate (Son et al., 2013). Again, the responsible genes are overexpressed in adenomas and cancer (*GOT2*, *MDH2*, *ME1*). In summary, these data indicate that many of the known metabolic changes, which are mostly induced by an upregulation of relevant enzyme genes, are observable already in premalignant colorectal lesions. Consequently, the metabolic adaptation within these pathways might be already an early event of colorectal carcinogenesis and persists during the progression to cancer. Interestingly, we see also many changes in the metabolome within these pathways. For example proline, glutamine, glutamate, 5-phosphoribosylamine, phosphoserine and glycine are increased. This gives us further evidence to believe that the biosynthetic activity for these substances is increased in early colorectal lesions, i.e adenomas. Interestingly however, most of the metabolites associated with these pathways were increased only in cancer and not in adenomas.

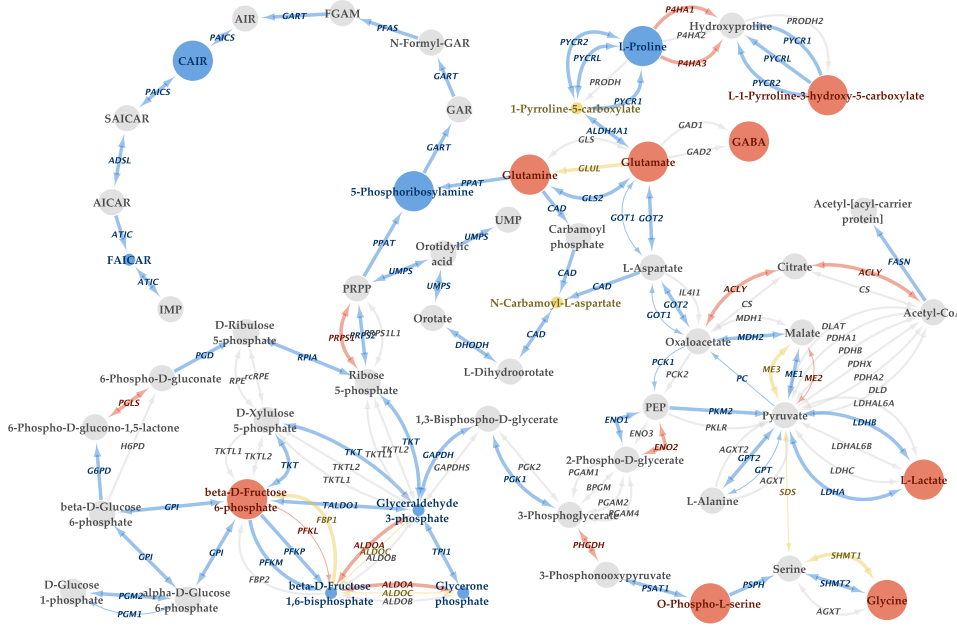


Figure 3.6: Metabolomic and transcriptomic changes of core metabolism in colorectal adenomas and cancer. Metabolites are represented as circles and enzymatic reactions are shown as arrows. Relevant genes for each enzymatic reaction are indicated on the arrows. Metabolic reactions were obtained from KEGG database (Ogata et al., 1999), the network was created using cytoscape (Barupal et al., 2012). Gene overexpression or metabolite increase is indicated in bold, gene downregulation or metabolite decrease is indicated in light print. Genes and metabolites are colored according to the BGA groups: blue: deregulation in adenoma and cancer, yellow: deregulation in adenoma, red: deregulation in cancer. Genes and metabolites without expression alteration or which could not be detected in the LCMS data are shown in grey.

3.4.4 Investigation of solute carrier transporters (SLC) and its substrates

Increased metabolism of glucose and glutamine is accompanied by an increased uptake of these substances into cancer cells (Kroemer and Pouyssegur, 2008). For example, the need for glucose is maintained by an increased production of the glucose transporter GLUT-1, a cancer hallmark which is nowadays exploited as a standard tool in clinical cancer diagnostics (Som et al., 1980; Kunkel et al., 2003). Uptake of metabolites into cells is mainly mediated through the so called solute carrier (SLC) transporter class (Hediger et al., 2013). Until now there are 52 transporter classes with a total of 395 genes described and maintained within the bioparadigms database (http://slc.bioparadigms.org). For many of these genes in the database there is a list of known transported subtances available. In order to further characterize the metabolic alterations of colorectal cancer and adenoma, we investigated the expression of SLC genes in our dataset (Figure 3.7). Of all the annotated 395 genes we could identify 376, from which 175 were classified according to the BGA analysis (93 deregulated in PCA, 58 in CA and 24 in P). Consistent with the expression of enzyme genes, most of the differentially expressed transporter genes are altered in adenoma and cancer, suggesting that a major part of metabolism is similar in adenomas and cancer. Of the upregulated genes within this class, many are involved in amino acid uptake, especially those of cationic nature (*SLC7A11*, *SLC6A14*, *SLC7A1*, *SLC7A6*, *SLC25A15*, Figure 3.7 a). The transporter gene *SLC6A6* with the highest fold change in this class is coding for a taurine transport protein. Interestingly, taurine itself showed up as one of the most increased compounds in our metabolomics dataset, also in adenoma and cancer (See Figure 3.3 a). For the cancer specific expression of transporter genes we observe neutral amino acid transporter genes (*SLC7A5, SLC3A2, SLC1A5*) and hexose transporter genes (*SLC2A3, SLC2A1*) as top 5 overexpressed genes (Figure 3.7 b). *SLC2A1* and *SLC2A3* are better known as *GLUT1* and *GLUT3* and known for their role as glucose transporter in cancer (Szablewski, 2013). *SLC7A5* (LAT1) and *SLC1A5* (ASCT2) are believed to work in a cooperative manner with the aim to import glutamine and other neutral amino acids (Fuchs and Bode, 2005).

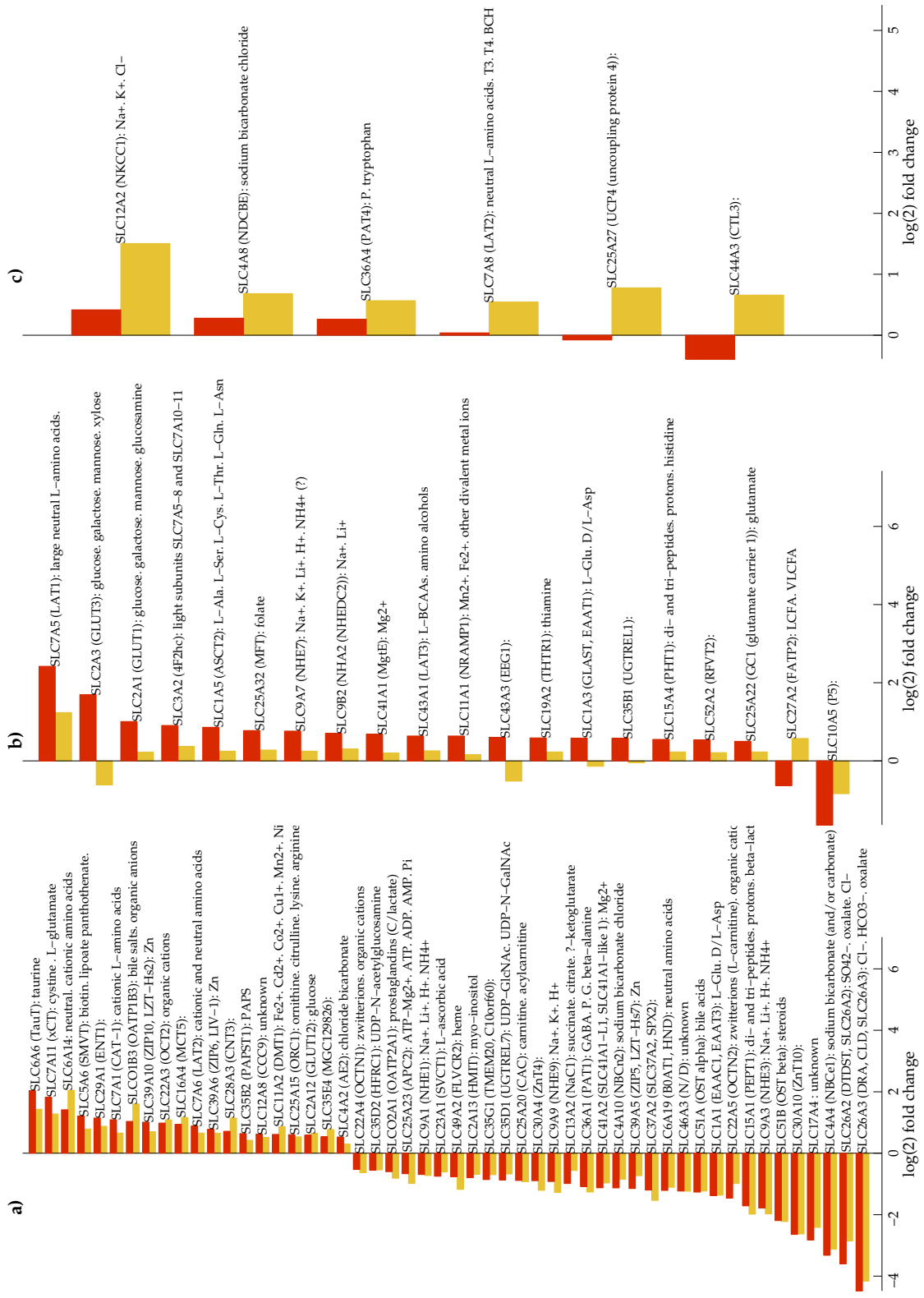


Figure 3.7: Log(2) fold change of SLC transporter gene expression for cancer versus normal mucosa (red) and adenoma versus normal mucosa (yellow). The expression values are sorted according to the maximal fold change. The columns were selected after the BGA classification: a) deregulated in adenoma and cancer, b) deregulated in adenoma, c) deregulated in cancer.

3.4.5 Upregulation of the *MYC* gene and its downstream targets in colorectal cancer and adenomas

Mutations in the APC gene and a subsequently increased expression of the *MYC* gene have been considered as key event in sporadic and hereditary initiation of colorectal tumors (Sansom et al., 2007). *MYC* itself is described as a transcription factor involved in the activation of multiple metabolic genes involved in glycolysis, glutamine, serine or proline metabolism (Li and Simon, 2013). Many of these genes are shown in this study to be upregulated in adenomas and cancer, resulting in an increase of metabolites within relevant metabolic pathways (Figure 3.6). We therefore hypothesized whether the activation of *MYC* in early stages of colorectal tumors is responsible for a metabolic reprogramming that persists during each stage of tumor progression. In order to test this hypothesis our idea was to investigate the expression of *MYC* and *MYC*-target genes in our dataset. We used a list of genes that were shown to be specific *MYC* targets independent of cell type (Table A-3) (Ji et al., 2011). 77 of these 98 genes were significantly deregulated in our transcriptomics dataset according to the BGA clustering. Of these, 92.2% as well as the *MYC* gene itself were classified as upregulated in cancer and adenomas (Figure 3.8). These data together indicate that many enzyme genes which are equally upregulated in adenomas and cancer might be regulated by the *MYC* transcription factor.

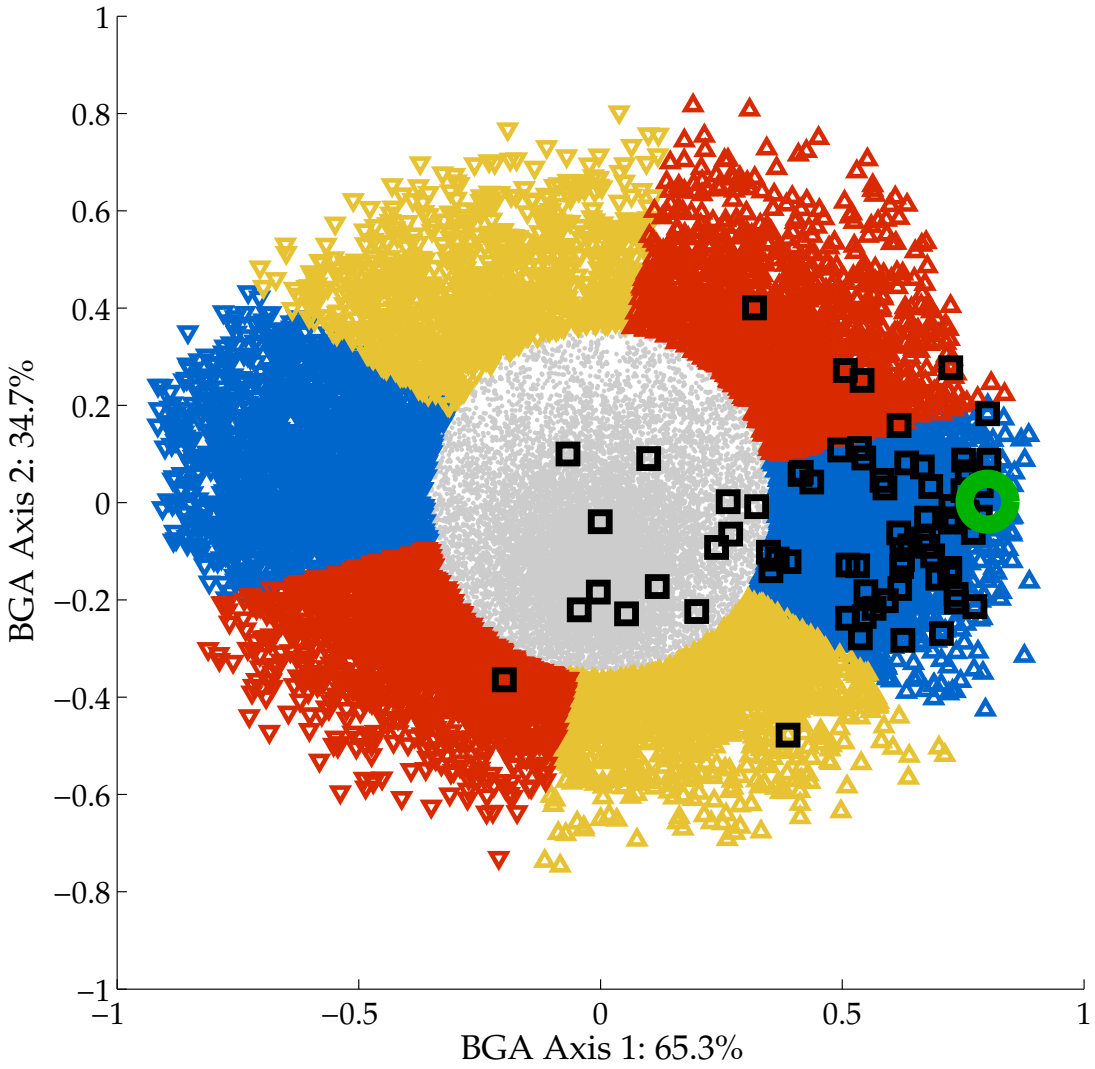


Figure 3.8: Loadings plot of the between group analysis (BGA) based on transcriptomics data. This is the same representation as seen in Figure 3.2 c). Green circle indicates position of the *MYC* gene, black boxes indicate positions of MYC target genes (See Table A-3).

3.4.6 Lipidomics profiling reveals a role for polyunsaturated, very long chained fatty acid synthesis in colorectal cancer development

Several pathways involved in lipid metabolism were shown to be significantly altered in our overrepresentation analysis (See Table 3.2). Interestingly, among the few adenoma specific metabolite changes were two very long chain polyunsaturated fatty acids (VLC-PUFAs), arachidonic acid and eicosatrienoic acid. We therefore wanted to investigate the lipidome composition of colorectal adenomas and cancer in more detail. Compared to the metabolomics and transcriptomics dataset, the variance within the lipidomics data was high and the NM, P and CA group were not well separated in the BGA analysis (Figure 3.1). Especially the polyp and normal mucosa groups showed a clear overlap. Still, a total number of 252 lipids were

significantly altered according to the BGA analysis. The predominant lipid classes were triradylglycerols (TG), followed by glycerophophocholines (PC) and glycerophosphoserines (PS) (Figure 3.9).

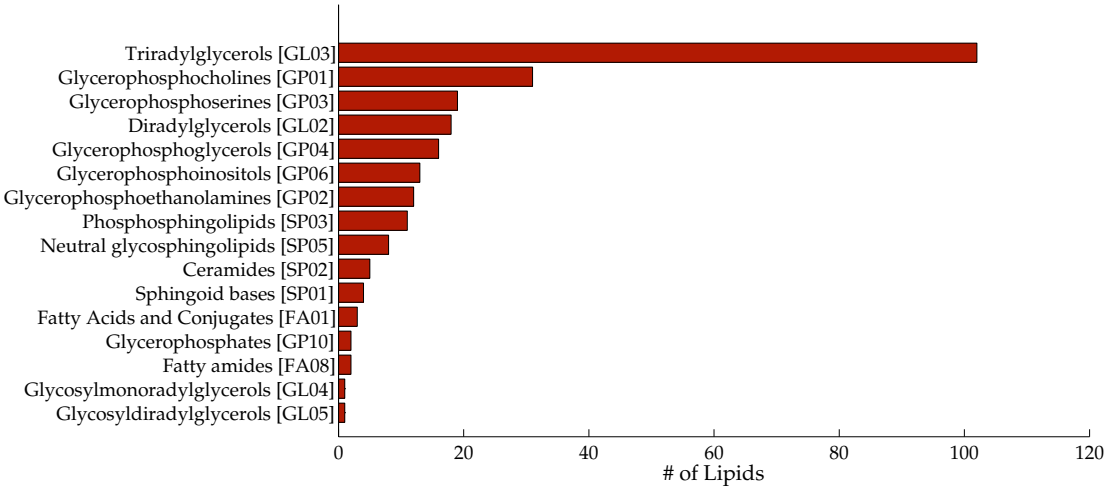


Figure 3.9: Distribution of deregulated lipids according to their lipidmaps classification (Fahy et al., 2009).

For each annotated lipid, we calculated the average carbon number (#C) and average double bond number (#n) of the attached fatty acyl residues (Figure 3.10). Most strikingly, almost all TG lipids with fatty acyl residues greater than 18 #C were upregulated in adenoma and cancer. About half of these fatty acyls showed a #n higher than 2, suggesting that most of the upregulated TGs contain VLC-PUFAs. In contrast, all decreased TGs contain fatty acyl groups with maximal 18 #C and less than 2 #n. The same trend was observed for Diradylglycerols (DG) and Glycerophosphoglycerols (PG) where also VLC-PUFA containing lipids were increased, and lipids with smaller and more saturated fatty acyls were decreased. In other lipid classes we observed different trends. Glycerophosphoinositols (GPI) for example were only found to be downregulated. This was true for GPI containing VLC-PUFA but also for shorter fatty acyl chains with less double bonds. In the glycerophophocholine (PC) class VLC-PUFA containing lipids were decreased and PCs with very long chain saturated fatty acyls were increased. Interestingly, the amount of fold change between lesion and normal mucosa was higher in cancer for many of the VLC-PUFA containing lipids, especially in the DG and PG class.

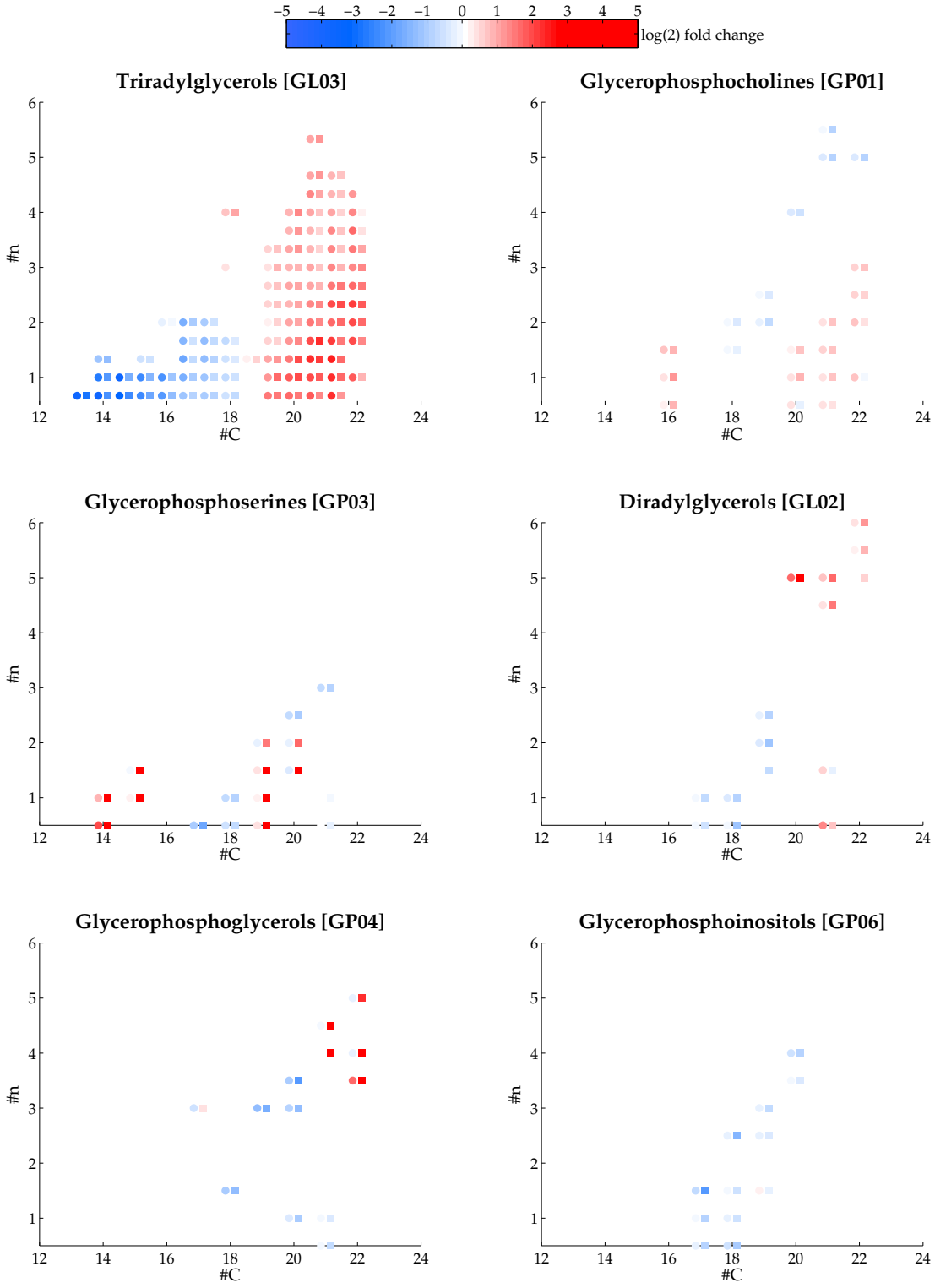


Figure 3.10: Average number of fatty acyl carbon (#C) and number of double bonds (#n) for each annotated lipid. The log(2) fold change is displayed for adenoma versus normal mucosa (circle) and cancer versus normal mucosa (box). Only the most 6 abundant lipid classes (Figure 3.9) are shown.

3.4.7 Expression of enzymes involved in lipid metabolism

The increase of different lipid classes with VLC-PUFA fatty acyl chains raised the question for an enzymatic mechanism which favours such long and unsatured fatty acyls in lipids. Some genes involved in fatty acid biosynthesis and fatty acid elongation were over represented in our pathway analysis. We therefore speculate that an increase in these pathways may lead to the phenotype with a high amount of VLC-PUFA in cancer and adenoma lipids. We had a detailed look at the genes involved in fatty acid metabolism pathway in KEGG, which includes fatty acid biosynthesis, elongation, desaturation and mitochondrial fatty acid metabolism. 36 of a total of 48 genes in this pathway were unambiguously classified according to the BGA analysis (See Figure 3.2). A detailed view of all reaction steps including the mapping of deregulated genes within the different BGA classes can be seen in Appendix Figure A-5. We summarized these steps according to the cellular location of the reactions (mitochondrial, cytoplasmic, endoplasmic reticulum ER) in Figure 3.11. Among the increased genes in adenoma and cancer are *FASN* and *ACACA*. These genes are important for the initiation of fatty acid synthesis in the cytoplasm and the generation of palmitate. Interestingly, most of the genes involved in fatty acid elongation or fatty acid desaturation are increased in cancer. Especially *FADS1*, *FADS2* and *SCD5* show exclusive upregulation in cancer whereas *SCD* is also increased in adenoma. Finally, almost all mitochondrial enzyme genes are downregulated. Many of these mitochondrial genes play a role in fatty acid oxidation (e.g. *ACOX1*, *CPT2*, *CPT1A*). Again, some of these genes are uniquely downregulated in cancer (*ACADM, ACADVL, ACADSB*), wheres others are deregulated in both adenoma in cancer. Together, these results imply that adenoma and predominantly cancer cells increase their de-novo biosynthesis, elongation and desaturation of fatty acids while genes involved in mitochondrial fatty acid metabolism are downregulated. As a complement to the lipidomics analysis, these gene alterations might explain the increase of VLC-PUFA containing lipids in the colorectal lesions, especially in cancer tissue.

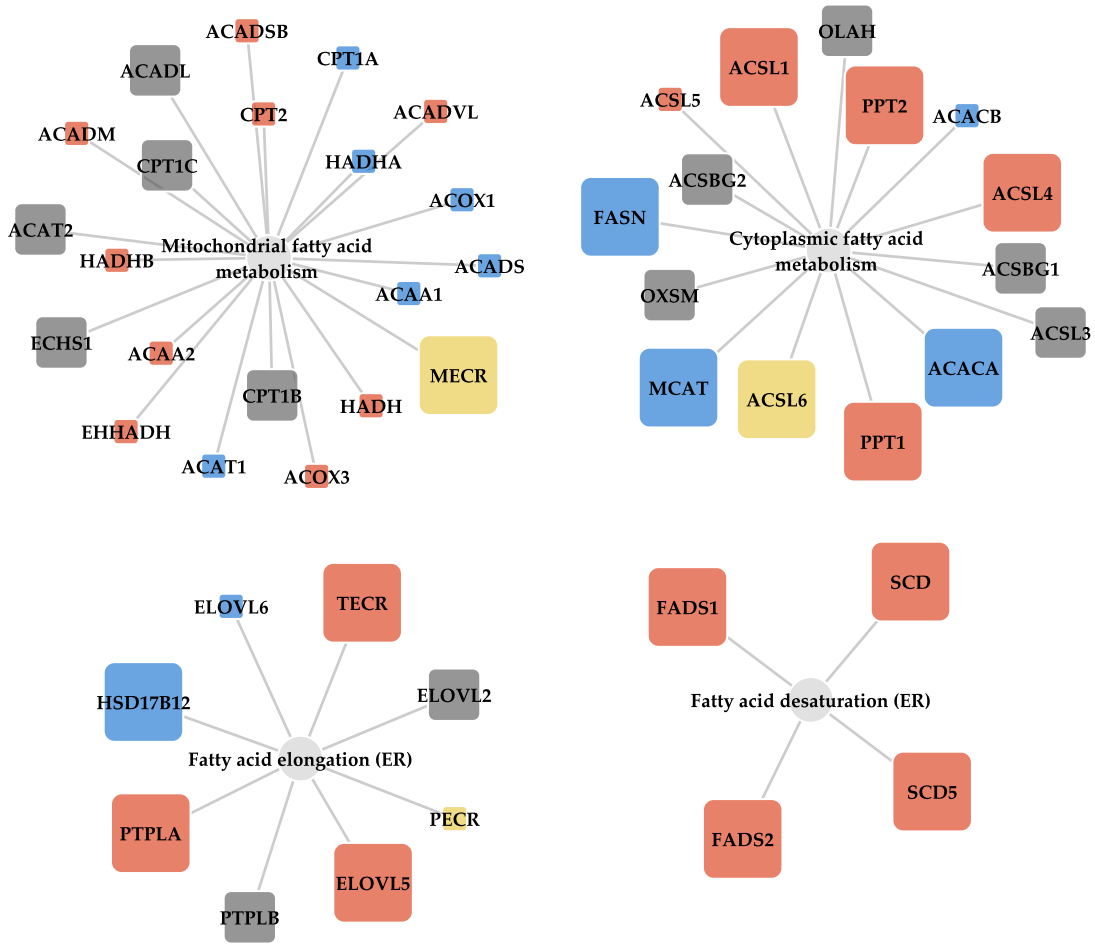


Figure 3.11: Deregulation of genes involved in fatty acid metabolism. Each gene is represented as box and is connected via an edge to its biological function displayed in grey. Gene overexpression is indicated in bold, downregulation in light print. Red: deregulated in cancer, blue: deregulated in cancer and adenoma, yellow: deregulated in adenoma.

3.5 Conclusion

In this study we investigated the metabolome and lipidome alterations of 49 colorectal lesions compared to adjacent mucosa from a total of 39 patients. For better interpretation of the metabolomics and lipidomics data we included transcriptomics data from our previous studies (Cattaneo et al., 2011; Maglietta et al., 2012). We applied a supervised multivariate method (between-group analysis, (Culhane et al., 2002)) to classify metabolites, lipids and transcripts according to their abundances in three tissue types: normal mucosa, adenoma and cancer. According to this classification most of the transcriptome, metabolome and lipidome alterations were similar in both lesion types, suggesting that most metabolic changes in colorectal cancer and

adenoma tissue are identical. We then investigated metabolic changes using pathway overrepresentation analysis and observed that central metabolic pathways such as glycolysis, nucleotide synthesis, proline synthesis, serine synthesis and fatty acid synthesis were affected at the metabolome and transcriptome level. Interestingly, most of the enzyme genes within these pathways are upregulated in both adenomas and cancer, suggesting that reprogramming of central metabolism is an early event in colorectal carcinogenesis. Previous studies in cancer metabolism have already shown that the *MYC* gene plays a central role in controlling central carbon metabolism in cancer (Li and Simon, 2013). Since upregulation of the *MYC* gene is one of the first events in colorectal tumor evolution (Sansom et al., 2007), we speculated about the role of *MYC* as one mediator of these metabolic changes. In support of this theory, we showed that the expression of *MYC* and 92.2% of a selection of *MYC* controlled genes (Ji et al., 2011) correlate with the expression of central carbon metabolism genes. Interestingly, many metabolite alterations were more dramatic in cancer than in adenomas such as for glutamate, glutamine or glucose. In addition, we observed solute carrier genes (*SLC*), which were overexpressed specifically in cancer. Among these *SLC* genes were transporter for glucose, glutamine and glutamate. Considering all these facts, we speculate that a specific overexpression of these transporters in cancer causes an increased intake of these metabolites compared to adenomas, which together might lead to a more aggressive proliferation. Two of these transporters, GLUT1 and LAT1, were already shown as predictive markers of poor response in clinical colorectal and rectal cancer therapy (Ebara et al., 2010; Shim et al., 2013). In addition to central carbon metabolism, fatty acid and lipid metabolism was altered on the transcriptome and metabolome level. Among the few adenoma specific metabolite alterations were two VLC-PUFA classes, arachidonic acid and eicosatrienoic acid. Lipidomics analysis revealed that adenomas and cancer showed an increase of lipids with VLC-PUFA fatty acyls while lipids with low chain, saturated fatty acyls were decreased. Along with these observations we found an increased gene expression for the biosynthesis and elongation of fatty acids, especially in cancer. Together, these results implicate de-novo synthesis and accumulation of VLC-PUFAs in colorectal adenomas and cancer. This synthesis capacity may play an important role in COX-2

related arachidonic acid metabolism of colorectal cancer (Oshima et al., 1996; Castellone, 2005).

Conclusions and perspectives

In the first two parts of this thesis we report the development of novel analytical strategies for untargeted metabolomics with the aim to improve overall sensitivity and coverage of different metabolite classes. We demonstrate two complementary capillary LCMS methods (capLCMS), one for the analysis of polar metabolites and one for unpolar metabolites and lipids. With both methods we achieve very good sensitivity in the femtomolar range and a broad coverage of the metabolome and lipidome from a low amount of colorectal tissue sample. Together, these methods enable us to cover about 50% of the masses annotated in the KEGG database and about 700 lipids in the lipidmaps database by accurate mass. With such an untargeted metabolomics approach we have the possibility to explore metabolic changes of multiple metabolic pathways. However, the identification of unknown metabolites in such an untargeted approach was mainly restricted to the annotation of accurate mass information to a metabolite database with known masses. The presence of isomeric compounds for example does not allow to unambiguously annotate each metabolite by one single mass. If the chromatographic dimension does allow for separation of isomers then additional information is needed to distinguish between them. In our case we demonstrate the separation of three hexose-phosphate classes using HILIC chromatography. For these compounds we could annotate the metabolite identity by retention time from reference standards. If no such standards can be used, annotation of each isomer by accurate mass alone is not possible. In order to improve the annotation of metabolites in an untargeted metabolomics experiment, mass fragment data can be included. This strategy not only may help to distinguish between isomers but also may confirm the identity of the metabolites which were annotated by accurate mass. Traditionally, data dependent acquisition (DDA) is used in exploratory mass spectrometry, especially in the area of proteomics (Mann et al., 2001). In this acquisition mode, mass spectrometers first generate a full scan mass spectrum, from which the top ten most intense ions are selected for subsequent MS/MS experiments. The selected ions, termed precursor ions, generate mass fragments which aid the identification of each precursor. Main disadvantages of this method are that only these high

abundant ions are considered while low abundant ions are ignored. Also, the duty cycle of DDA needs to be considered and consequently the chromatographic separation should not be too fast. In untargeted metabolomics however, mass spectra are typically recorded in MS full scan mode and further MS/MS experiments are performed to confirm the identity of a small list of interesting masses, either together with a pure standard or by parsing the MS/MS spectra to a database (Patti et al., 2012). This process is very time consuming and needs much additional manual work. An alternative to DDA is data independent acquisition (DIA). In this acquisition mode, two parallel alternating scans are recorded either with low or high collision energy (Plumb et al., 2006). In variation first implemented by Waters and termed MSE (MS with "energy"), DIA provides fragmentation of all ions acquired during a full scan MS and provides the potential to annotate fragment information for all ions. Typically, the duty cycle for these acquisition modes results in 50% low energy and 50% high collision energy time which makes this technique suitable for fast analysis with rapidly eluting chromatographic peaks. Especially in untargeted lipidomics applications, where many different isomeric compounds are expected, DIA was applied with good success (Castro-Perez et al., 2010; Hummel et al., 2011; Knittelfelder et al., 2014). Our analytical methods which were described in this thesis are designed to provide a fast and sensitive metabolome and lipidome analysis. The additional use of DIA techniques would further improve the interpretation and annotation of metabolites. For example, with the exact mass information we were able to annotate most lipids with a total amount of carbon and double bond in their fatty acyl chains. Therefore, the interpretation allows only for an average length and saturation status of the fatty acyl chains. Using DIA information it would be possible to identify the fatty acyl chain length as well as saturation status for each individual lipid (Castro-Perez et al., 2010). Since untargeted metabolomics strategies involve the detection of thousands of unknown masses, there is also a need for an automatic and comprehensive interpretation of such fragmentation data. In DDA, the precursor ions are isolated from disturbing background and the resulting fragments can be directly referred back to each precursor. In contrast, precursor-fragment assignment in DIA is challenging and until today there is no widely adopted software available for the automatic

processing of DIA data in untargeted metabolomics. Recently, a strategy using open source software tools was described to annotate precursor-fragment relationships (Broeckling et al., 2013). In this study, low collision and high collision energy MS data were separately analysed with XCMS and CAMERA in order to generate two independent peak lists. Using correlation analysis between the ion intensities of the low energy and the high energy peak list, the authors showed that precursor-fragment associations could be retrieved and the resulting spectra were highly similar to the results of a MS/MS experiment. In this thesis we described the development of the software tool cosmiq, which improves the detection of low abundant ion signals by combining multiple LCMS runs prior to peak detection. The implementation of the cosmiq peak detection pipeline in such a DIA workflow has the potential to increase the number of detected fragment ions and therefore would also improve the assignment of low abundant fragment ions to their potential precursors. Especially for the identification of fatty acyl chains in untargeted lipidomics, the relevant fragment intensities for lipids are very low (See Castro-Perez et al. 2010, Figure 7). In summary, the combination of our capillary LCMS methods and the software package cosmiq, in addition to a DIA workflow has the potential to further improve data interpretation of untargeted metabolomics and lipidomics experiments. An application example for the described analytical methods and the software package cosmiq (without DIA) was reported in the third chapter of this thesis. In a combined analysis of metabolomics, lipidomics and transcriptomics data we report an exploratory study of metabolic changes during colorectal cancer development. A basic finding was that in premalignant lesions and cancer tissue, many enzymes involved in central metabolism pathways were upregulated in addition to an increase of relevant metabolite within these pathways. Some of these enzymes were already known to be upregulated as part of metabolic reprogramming in cancer. Our data however suggest that a significant portion of metabolic reprogramming is already observable in premalignant lesions and therefore would explain an adaptation of proliferating cells in early stages of tumorigenesis. We also found metabolic alterations which were more prominent in cancer compared to the premalignant lesions. Among them is an increased amount of amino acids, an increased overexpression of SLC transporter for glucose and neutral amino

acids, and an increase in VLC-PUFA synthesis. As a general perspective, these exploratory data may serve as basis for deeper investigations of the observed alterations in colorectal cancer metabolism. One hypothesis we draw from this study is that pathways such as glycolysis and nucleotide synthesis are induced by an increase in enzyme activity in premalignant lesions and therefore early proliferating cells might have an activated anabolic metabolism for the synthesis of building blocks such as DNA, which is important for proliferation. However, several transporters for glucose and glutamine are only upregulated in cancer but not in adenomas. Glucose and glutamine are important fuels for the energetic and biosynthetic needs of the proliferating cells and we speculate that this increased uptake of glucose and glutamine might lead to a more aggressive proliferative phenotype of cancer. Future experiments might test this hypothesis, for example by generating two model systems, where the *MYC* gene and one or more of the cancer specific transporter genes such as *SLC7A5, SLC2A3* or *SLC2A1* are induced. Also, our data and methods might be used as the basis for biomarker discovery studies. It has already been shown that metabolites are altered in the urine or plasma of patients with cancer and can distinguish between healthy and diseased subjects with promising specificity and sensitivity (Farshidfar et al., 2012; Li et al., 2012; Manna et al., 2014). The development of adenoma specific biomarkers could be a clinical benefit for the early diagnosis of premalignant tumors. With the knowledge of the metabolic activities inside a tumor biopsy it is possible to make predictions about the alterations of metabolites which might be excreted from the tumor and are suspected to be found in body fluids. In addition to alterations in the metabolome it might be very helpful to consider the information about solute carrier transporter information into such predictions, especially those transporters which are important for the excretion of metabolites outside the cell. It is already known that cancer cells excrete lactate using the transporter MCT4 (Meijer et al., 2012). The knowledge about SLC overexpression and their relevant metabolite substrates, together with the observed alterations in the metabolome might help to generate hypotheses to find novel diagnostic opportunities to detect premalignant lesions in body fluids.

Appendix

Table A- 1: Composition of internal standard mixtures. Each metabolite standard was dissolved in HPLC grade water. Each lipid standard was dissolved in 1:1 chloroform:methanol. Metabolite internal standards were obtained from Cortecnet, Voisins-Le-Bretonneux, France. Lipids were obtained from Avanti Polar Lipids (Alabaster, AL, USA).

Name	Label	Mass	Concentration
Metabolites:			
Isovaleric Acid-1-13C, 99% 13C	1 C13	103.03505	2 mM
SODIUM PYRUVATE-2-13C, 99% 13C	1 C13	89.0194	2 mM
Malonic Acid-13C3, 99% 13C	All C13	107.021025	2 mM
Taurine-15N	1 N15	126.011701	2 mM
D-GLUCOSE-1-13C, 99% 13C	1 C13	181.066745	2 mM
D-Fructose-1-13C	1 C13	181.066745	2 mM
AMP-13C10,15N5, 98% 13C, 96-98% 15N	98% 13C, 96-98% 15N	362.081813	2 mM
Dimethyl succinate-2,2,3,3-d4	4 D	150.083018	2 mM
Sodium L-Lactate-1-13C solution, 99% 13C	1 C13	91.03505	2 mM
D-Sorbitol-1-13C	1 C13	183.082395	2 mM
UREA-15N2, 98% 15N	All N15	62.026433	2 mM
Algal Amino Acid Mixture-13C-15N:	98% 13C, 98% 15N		2 mg/ml
Lipids:			
1,2-diheptadecanoyl-*sn*-glycero-3-phospho-(1'-*rac*-glycerol)		772.523	1 mg/ml
1,2-diheptadecanoyl-*sn*-glycero-3-phosphoethanolamine		719.547	1 mg/ml
1-heptadecanoyl-2-hydroxy-*sn*-glycero-3-phosphocholine		509.348	1 mg/ml
1,2-diheptadecanoyl-*sn*-glycero-3-phosphocholine		761.593	1 mg/ml
1,2,3-triheptadecanoyl-glycerol		848.78	1 mg/ml
Margaric acid		270.2559	1 mg/ml

Table A- 2: Composition of 89 metabolite mixture used for the chracterization of stationary phases. Each compound in the mix has a concentration at 100 uM.

2-Oxoglutarate / alphaKetoGlutarate	Galactose	Mannose
2,3 BisPhosphoGlycerate	gamma-amino-butyricacid	Methionine
3-methyl-2-oxobutyrate (KetoValin)	Glucosamine-6-phosphate	Mevalonolactone
3-methyl-2-oxovalerate (KetoIsoLeu)	Glucose	Myo inositol
3-Phosphoglycerate	Glucose 1-phosphate	N-acetyl-glucosamine
4-methyl-2-oxovalerate (KetoLeu)	Glucose 6-phosphate	N2-Acetyl Lysine
6-phosphogluconate	Glucuronate	Nicotinate
Adenine	Glutamine	Ornithine
Adenosine diphosphate	Glycerate	Oxaloacetate
Adenosine monophosphate	Glycerol	Pantothenate
Adenosine triphosphate	Glycerol-phosphate	Phenylalanine
Alanine	Glycine	Phenylpyruvate
alpha-amino-butyricacid	Glyoxylate	Phosphoenolpyruvate
Arginine	Guanine	Proline
Asparagine	Guanosine diphosphate	Pyruvate
Aspartate	Guanosine monophosphate	Ribose
Carbamoyl phosphate	Guanosine triphosphate*	Ribose 5-phosphate
Citrate	Histidine	Ribose-1-phosphate
cyclo-Adenosin monophosphate	Homoserine	Ribulose 5-phosphate
cyclo-Guanosine monophosphate	Hydroxy-proline	Serine
Cysteine	iso-Citrate	Shikimate
Cystine	Isoleucine	Succinate
Diamino-pimelate	L-Glutamic acid	Sucrose
Dihydroxyacetonephosphate	L-Histidinol	Threonine
Erythrose 4-phosphate	L-Malate	Trehalose
Fructose	Lactate	Tryptophane
Fructose 1-phosphat	Leucine	Tyrosine
Fructose 1,6-biphosphate	Lysine	V aline
Fructose 6-phosphate	Maltose	Xylose
Fumarate	Mannitol	

Table A- 3: List of 98 MYC target genes. This list was retrieved from Ji et al. (2011).

Entrez:ID	Gene	Entrez:ID	Gene	Entrez:ID	Gene	Entrez:ID	Gene	Entrez:ID	Gene
6723	SRM	4691	NCL	91942	NDUFA12L	6418	SET	81892	C14orf156
26135	SERBP1	51540	SCLY	55781	RIOK2	6838	SURF6	55055	ZWILCH
26135	SERBP1	6599	SMARCC1	51520	LARS	55847	ZCD1	51728	POLR3K
26135	SERBP1	26354	GNL3	10146	G3BP1	9188	DDX21	55720	TSR1
26135	SERBP1	26354	GNL3	10146	G3BP1	56652	PEO1	27043	PELP1
10885	WDR3	26354	GNL3	4869	NPM1	23082	PPRC1	5198	PFAS
7203	CCT3	8607	RUVBL1	4869	NPM1	9221	NOLC1	57003	CCDC47
7203	CCT3	6434	SFRS10	4869	NPM1	54663	WDR74	25926	NOL11
7203	CCT3	5471	PPAT	51491	HSPC111	7004	TEAD4	3609	ILF3
54865	GPATC4	10606	PAICS	7919	BAT1	7004	TEAD4	2091	FBL
54865	GPATC4	55153	SDAD1	7919	BAT1	7004	TEAD4	27338	UBE2S
1660	DHX9	3184	HNRPD	3326	HSP90AB1	58516	FAM60A	2618	GART
51514	DTL	3184	HNRPD	134637	DEADC1	1594	CYP27B1	2618	GART
57539	WDR35	3184	HNRPD	7532	YWHAG	4234	METTL1	23481	PES1
57539	WDR35	3184	HNRPD	54984	PINX1	4234	METTL1	4809	NHP2L1
64682	ANAPC1	9987	HNRPDL	25879	WDSOF1	29902	C12orf24	4809	NHP2L1
84172	POLR1B	54433	NOLA1	65083	NOL6	328	APEX1	92745	SLC38A5
79828	METTL8	54433	NOLA1	65083	NOL6	328	APEX1	1736	DKC1
84128	WDR75	79960	PHF17	55035	NOL8	328	APEX1		
51602	NOP5/NOP58	6059	ABCE1	54107	POLE3	11198	SUPT16H		

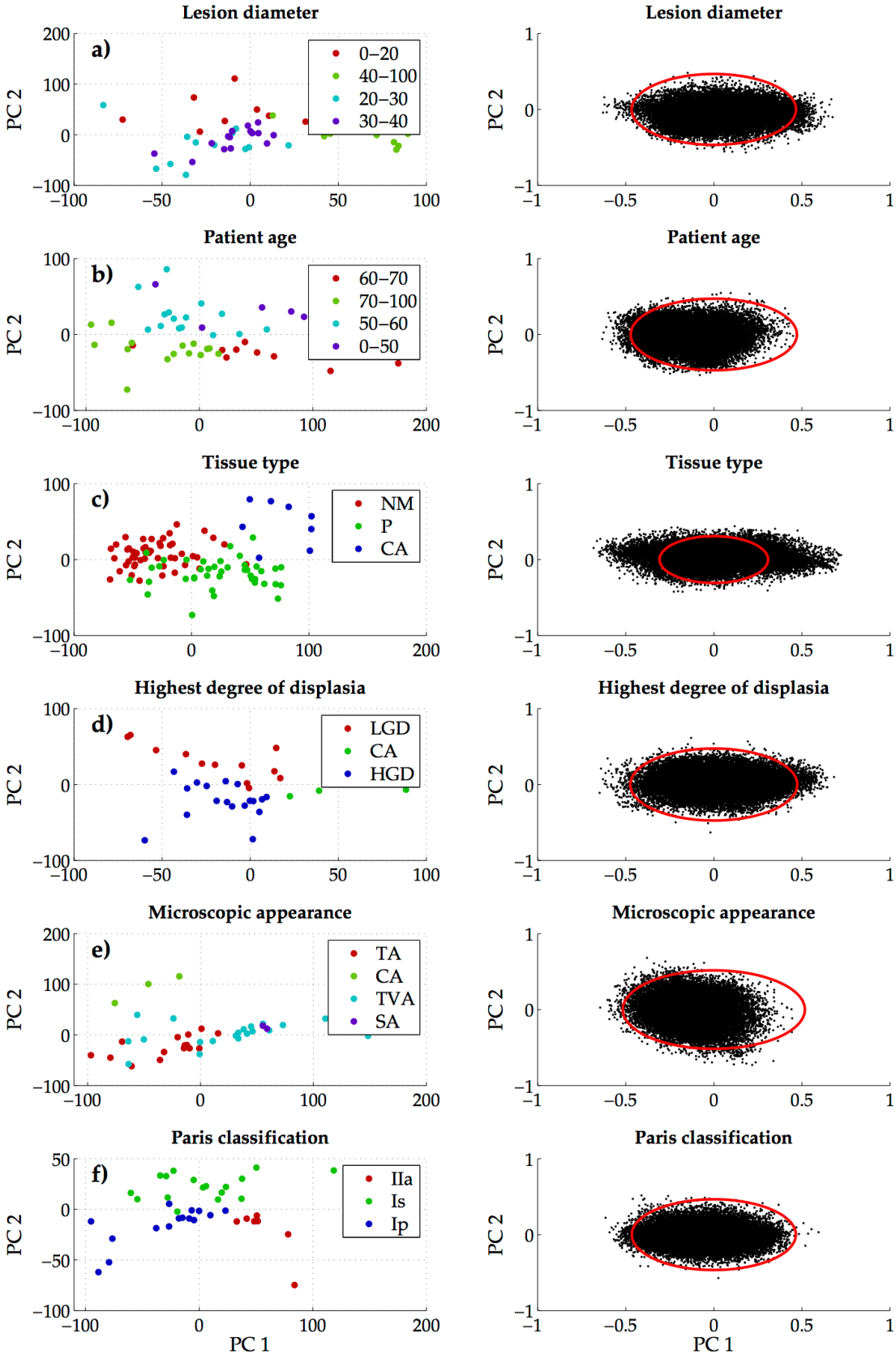


Figure A- 1: Between group analysis (BGA) of colorectal adenoma metabolome based on different clinical parameters (a-f) as grouping information. Left panel shows the score plots of the BGA analysis and the loadings (metabolites) oft he same analysis are shown in the right panel. The red ellipse indicates the 99% percentile of randomly generated loadings after a permutation test (See materials and methods). Loadings outside the ellipse are considered as significant.

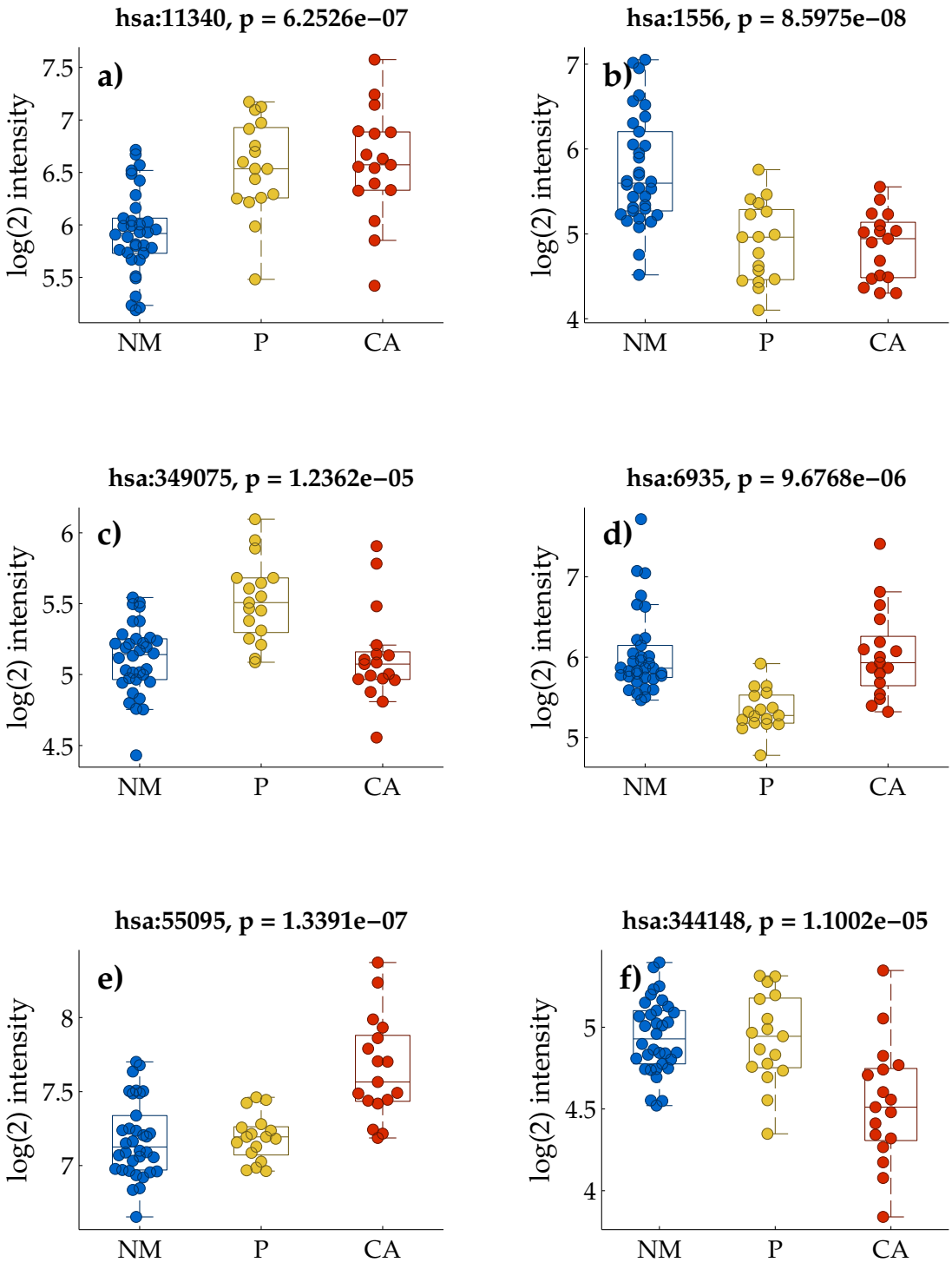


Figure A- 2: Expression values of selected genes based on the BGA classification (see Figure 3.2 c). The genes were randomly selected from the center of each classified group in the scatter plot: a) Up in P and CA, b) Down in P and CA, c) up in P, d) down in P, e) up in CA, f) down in CA. NM = normal mucosa, P = polyp, CA = cancer. The entrez gene name is shown. For each gene, a p-value was obtained by one-way anova analysis using NM, P and CA as group information.

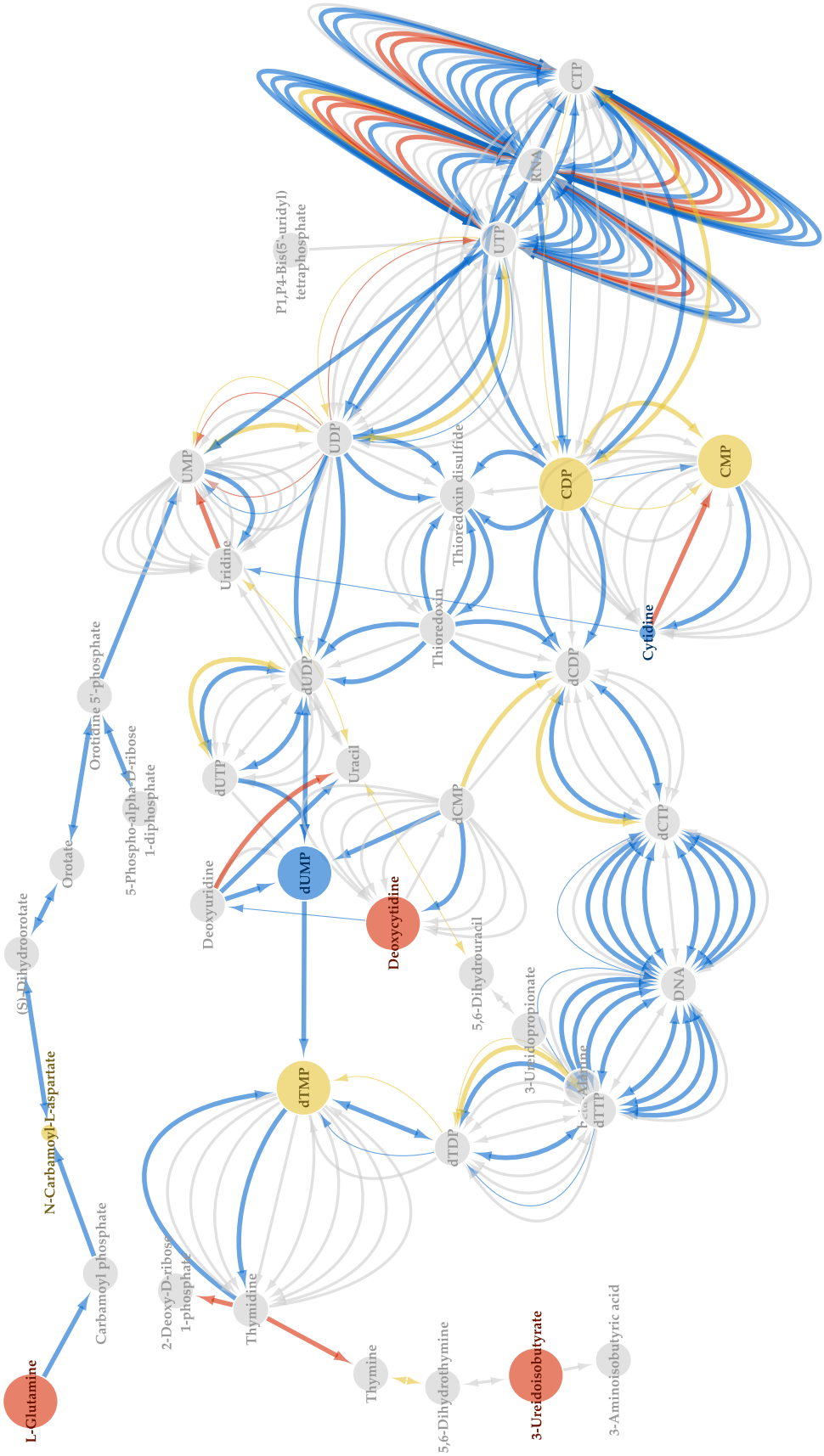


Figure A- 3: Pyrimidine metabolism pathway. The pathway map was generated from the KEGG kgml data files (Ogata et al., 1999) using cytoscape (Shannon et al., 2003). The annotation is the same as in Figure 3.6. The gene label is not displayed for the purpose of better visibility.

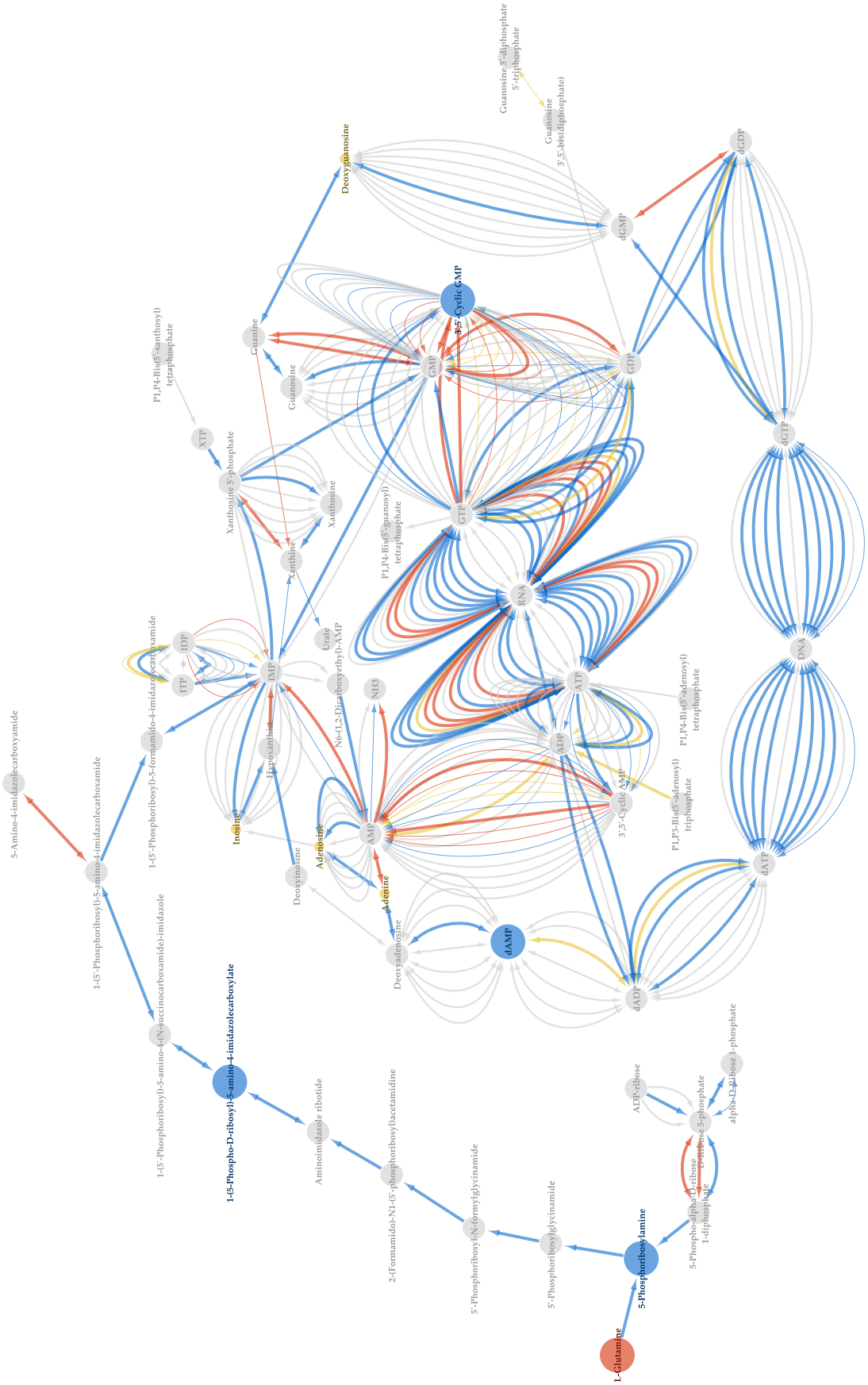


Figure A- 4: Purine metabolism pathway. This Figure was created as explained in Figure A-3.

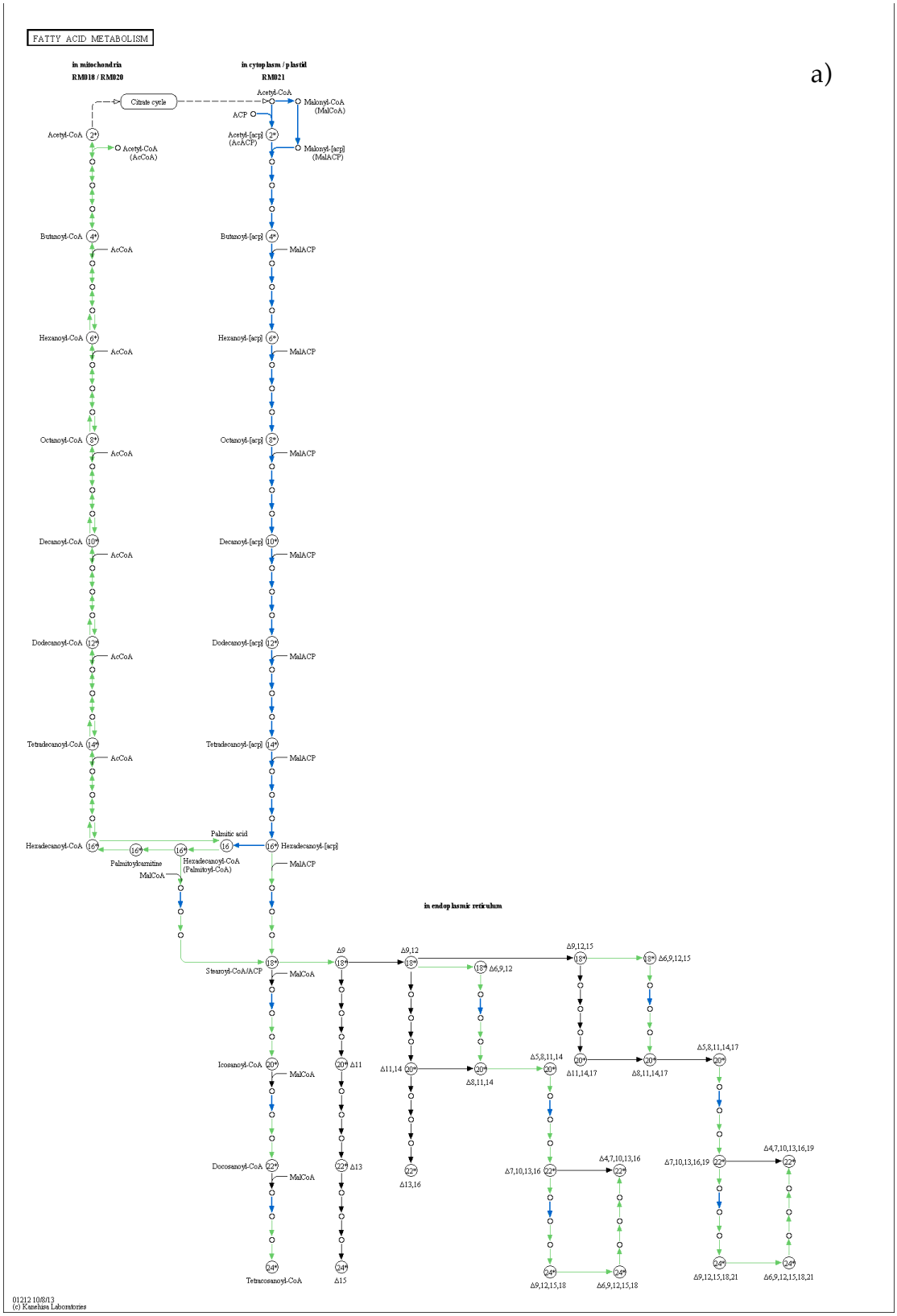


Figure A- 5: Mapping of enzyme gene expression changes to KEGG pathway map "fatty acid metabolism". Each arrow represents one reaction step. Bright green: not affected human gene. Blue: Deregulated in adenoma (P) and cancer (CA). Yellow: Deregulated in adenoma (P). Red: Deregulated in cancer (CA). a) Upregulated in PCA, b) downregulated in PCA, c) upregulated in P, d) downregulated in P, e) upregulated in CA, f) downregulated in CA.

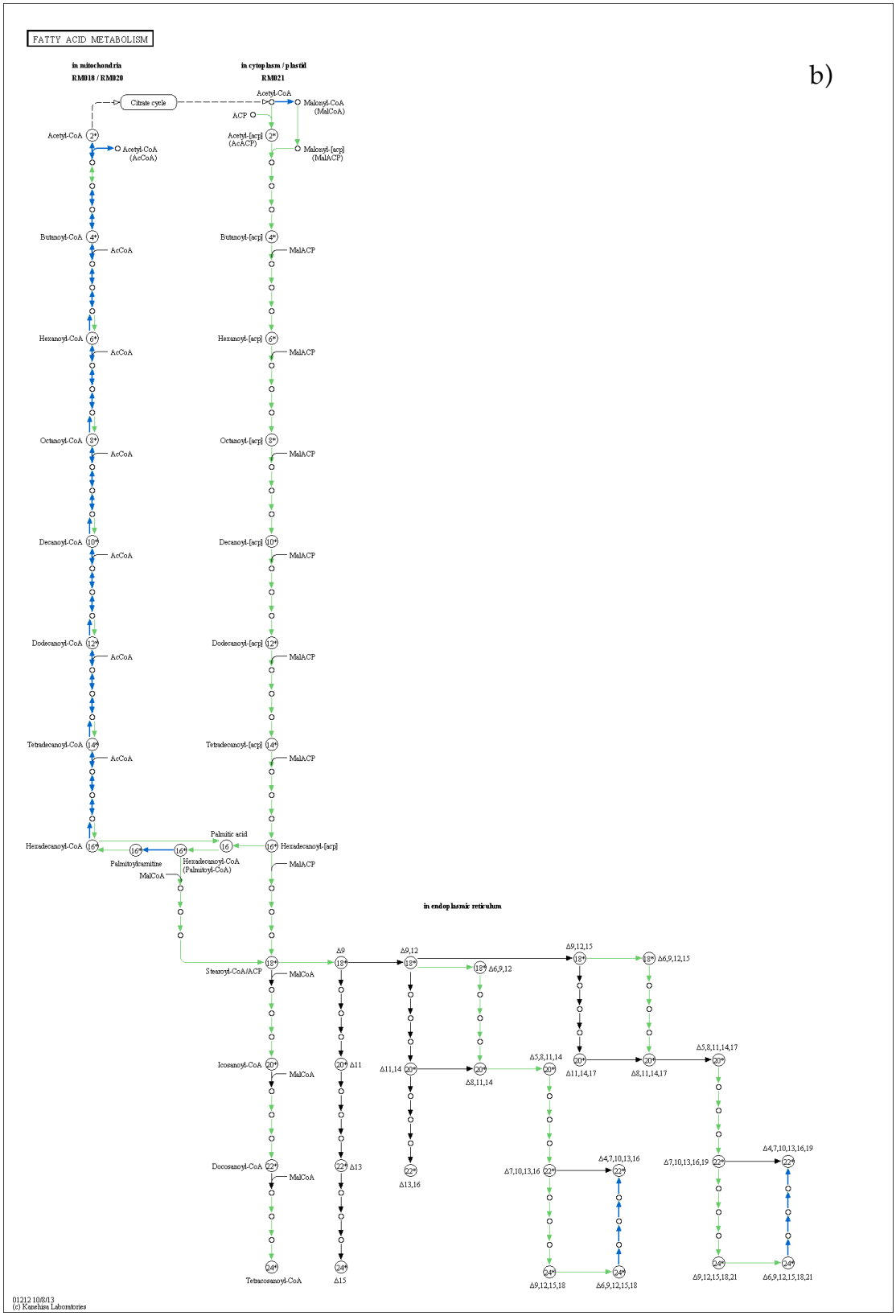


Figure A- 5 (continued)

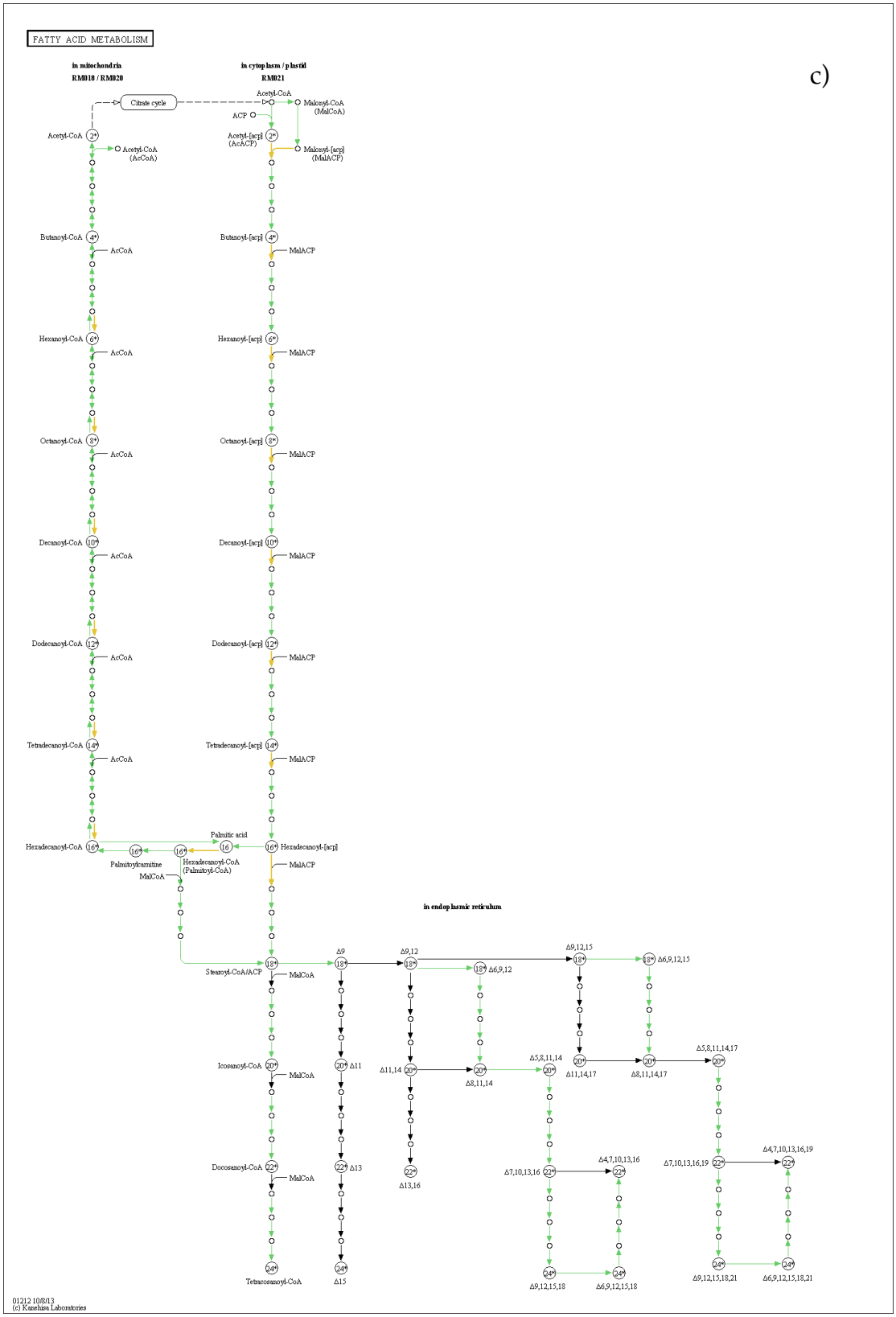


Figure A- 5 (continued)

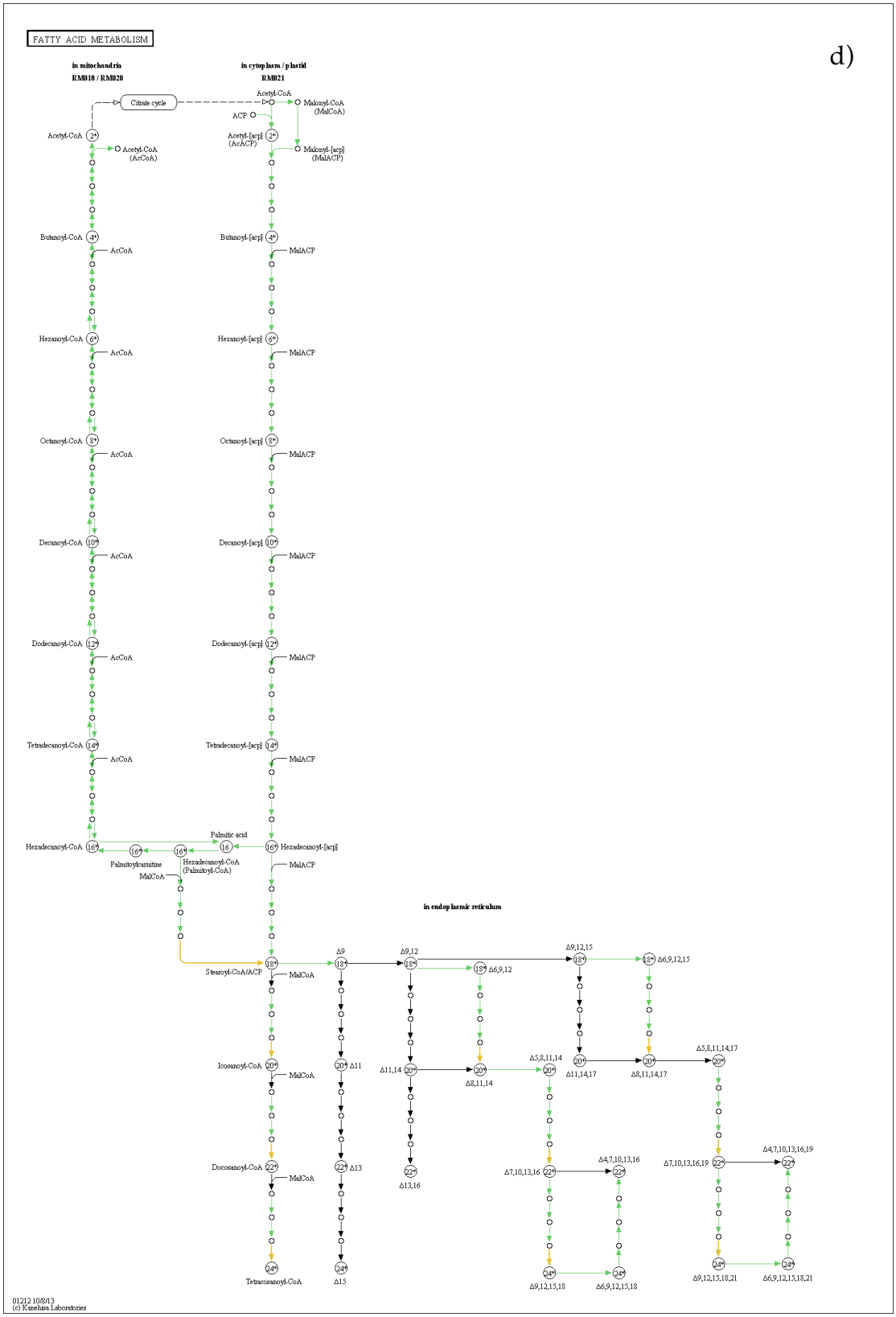


Figure A- 5 (continued)

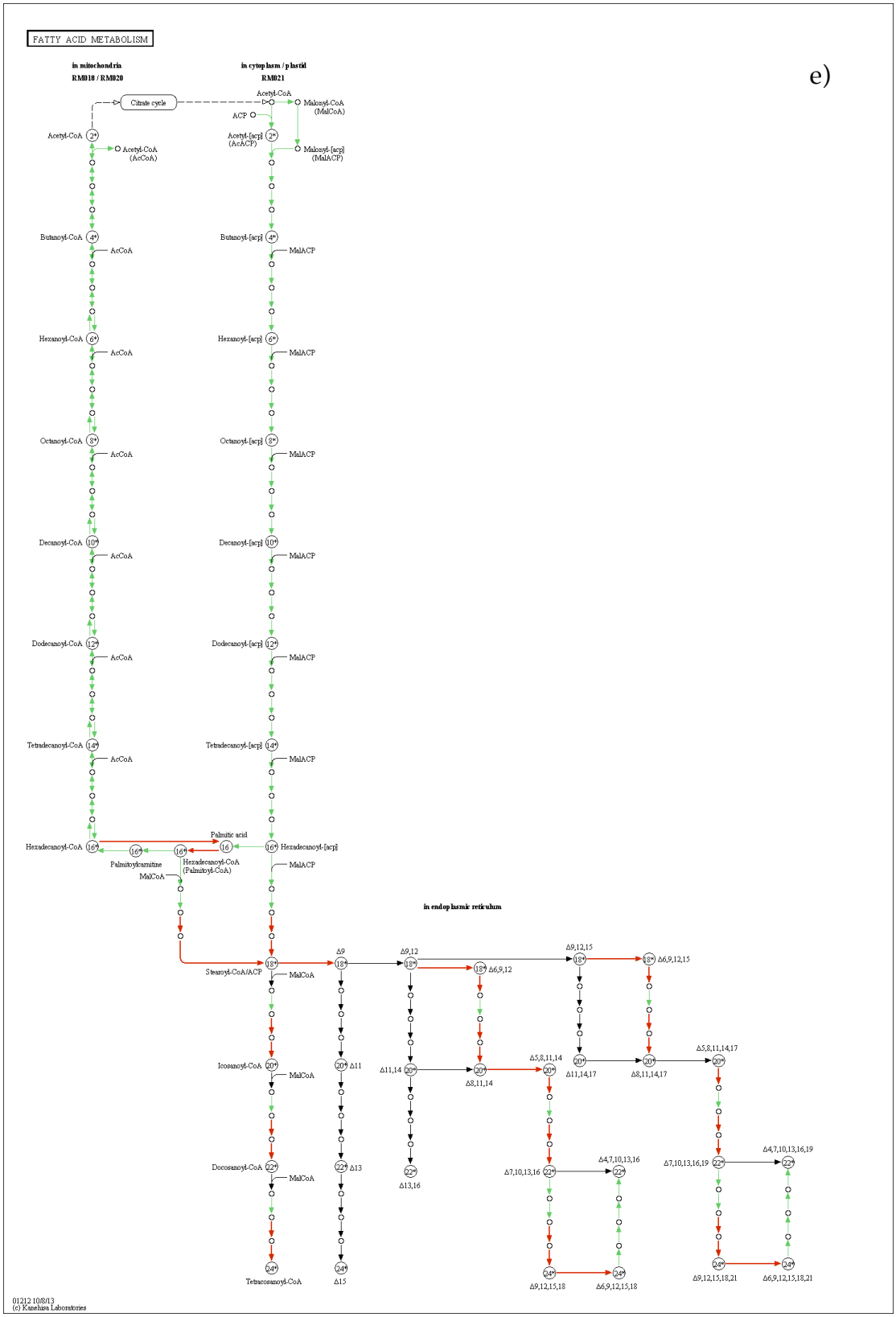


Figure A- 5 (continued)

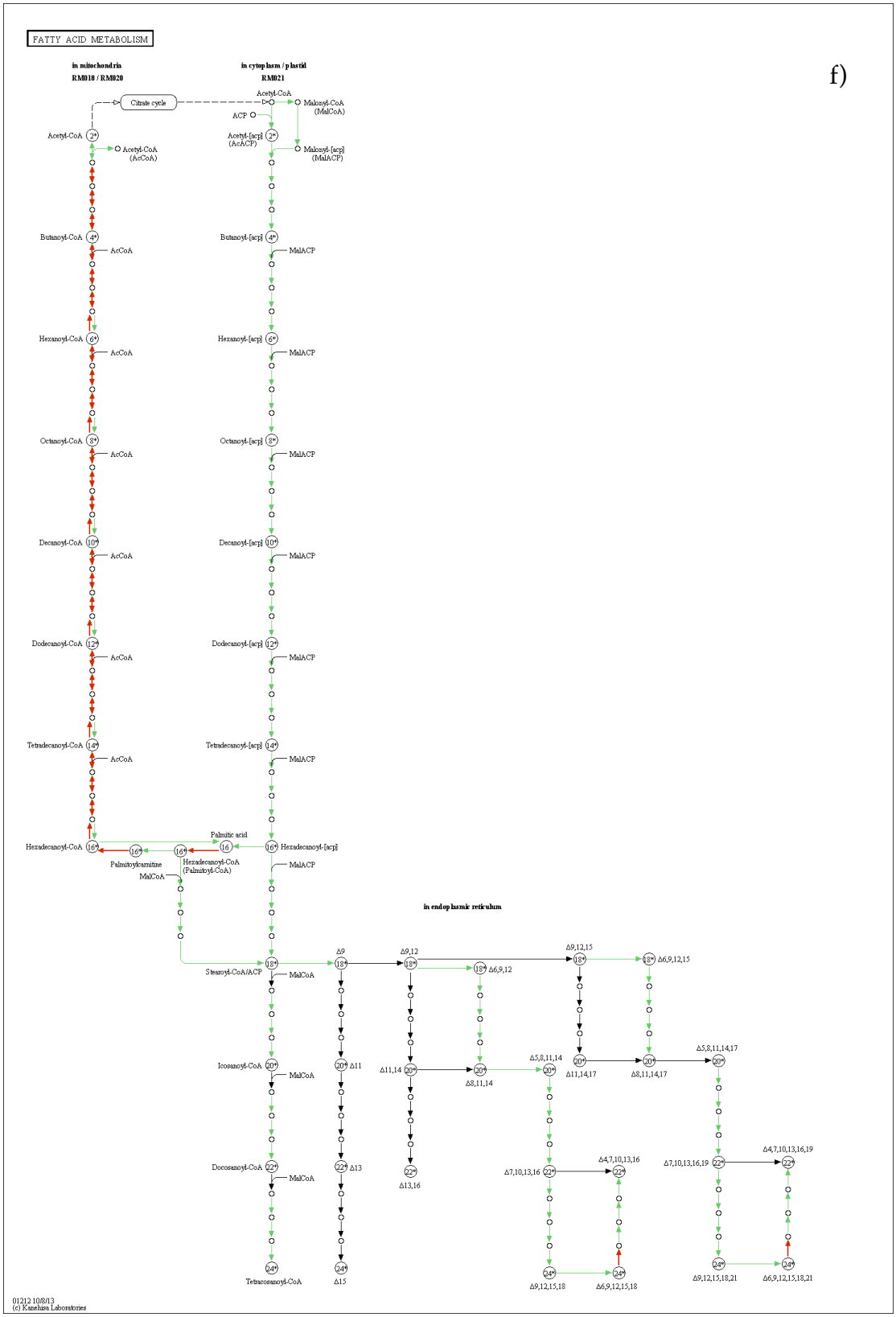


Figure A- 5 (continued)

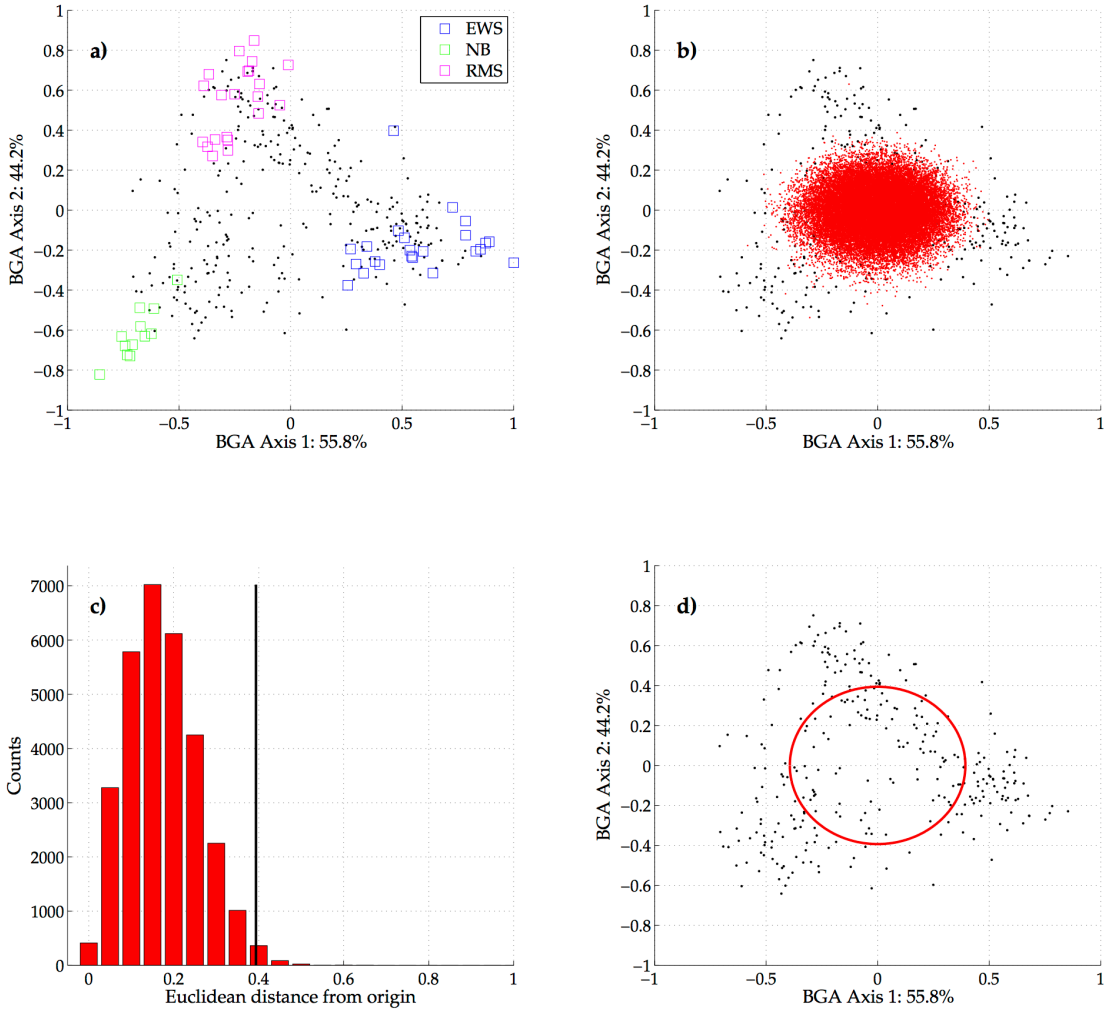


Figure A- 6: Workflow to select significant variables from different omics datasets based on BGA and a permutation test, demonstrated on an example transcriptomics dataset (Golub et al., 1999). a) Biplot of the BGA results are shown with loadings in black and scores in color (EWS=Ewing family of tumours, NB = neuroblastoma, RMS = rhabdomyosarcoma). b) loadings plot of the same dataset (black) and loadings of 100 permutations of this data. c) histogram of the vector length between origin and BGA Axis 1 and 2 of all permuted loadings. Black line indicates the 99% percentile of these vector lengths. d) scatter plot of the original data as seen in a) and b) in black, showing additionally the 99% percentile of the randomly permuted loading vector lengths as radius of a red circle. All the loadings inside this circle are considered as randomly generated with 99% confidence.

References

Aberg, K.M., Torgrip, R.J., Kolmert, J., Schuppe-Koistinen, I., and Lindberg, J. (2008). Feature detection and alignment of hyphenated chromatographic–mass spectrometric data. Journal of Chromatography A *1192*: 139–146.

Abian, J., Oosterkamp, A.J., and Gelpí, E. (1999). Comparison of conventional, narrow-bore and capillary liquid chromatography/mass spectrometry for electrospray ionization mass spectrometry: practical considerations. J. Mass Spectrom. *34*: 244–254.

Bajad, S.U., Lu, W.Y., Kimball, E.H., Yuan, J., Peterson, C., and Rabinowitz, J.D. (2006). Separation and quantitation of water soluble cellular metabolites by hydrophilic interaction chromatography-tandem mass spectrometry. Journal of Chromatography A *1125*: 76–88.

Baker, M. (2011). Metabolomics: from small molecules to big ideas. Nat Methods *8*: 117–121.

Barker, N., Ridgway, R.A., van Es, J.H., van de Wetering, M., Begthel, H., van den Born, M., et al. (2008). Crypt stem cells as the cells-of-origin of intestinal cancer. Nature *457*: 608–611.

Barupal, D.K., Haldiya, P.K., Wohlgemuth, G., Kind, T., Kothari, S.L., Pinkerton, K.E., et al. (2012). MetaMapp: mapping and visualizing metabolomic data by integrating information from biochemical pathways and chemical and mass spectral similarity. BMC Bioinformatics *13*: 99–2504.

Baysal, B.E., Ferrell, R.E., Willett-Brozick, J.E., Lawrence, E.C., Myssiorek, D., Bosch, A., et al. (2000). Mutations in SDHD, a Mitochondrial Complex II Gene, in Hereditary Paraganglioma. Science *287*: 848–851.

Benjamin, D.I., Cravatt, B.F., and Nomura, D.K. (2012). Global profiling strategies for mapping dysregulated metabolic pathways in cancer. Cell Metabolism *16*: 565–577.

Bensaad, K., Tsuruta, A., Selak, M.A., Vidal, M.N.C., Nakano, K., Bartrons, R., et al. (2006). TIGAR, a p53-Inducible Regulator of Glycolysis and Apoptosis. Cell *126*: 107–120.

Broeckling, C.D., Heuberger, A.L., Prince, J.A., and Ingelsson, E. (2013). Assigning precursor–product ion relationships in indiscriminant MS/MS data from non-targeted metabolite profiling studies. Metabolomics *9*: 33–43.

Buescher, J.M., Moco, S., Sauer, U., and Zamboni, N. (2010). Ultrahigh Performance Liquid Chromatography−Tandem Mass Spectrometry Method for Fast and Robust Quantification of Anionic and Aromatic Metabolites. Anal. Chem. *82*: 4403–4412.

Castellone, M.D. (2005). Prostaglandin E2 promotes colon cancer cell growth through a Gs-axin-beta-catenin signaling axis. Science *310*: 1504–1510.

Castro-Perez, J.M., Kamphorst, J., DeGroot, J., Lafeber, F., Goshawk, J., Yu, K., et al. (2010). Comprehensive LC−MS ELipidomic Analysis using a Shotgun Approach and Its Application to Biomarker Detection and Identification in Osteoarthritis Patients. J. Proteome Res. *9*: 2377–2389.

Cattaneo, E., Baudis, M., Buffoli, F., Bianco, M.A., Zorzi, F., and Marra, G. (2010). Pathways and Crossroads to Colorectal Cancer. In *Pre-Invasive Disease: Pathogenesis and Clinical Management*, (New York, NY), pp 369–394.

Cattaneo, E., Laczko, E., Buffoli, F., Zorzi, F., Bianco, M.A., Menigatti, M., et al. (2011). Preinvasive colorectal lesion transcriptomes correlate with endoscopic morphology (polypoid vs. nonpolypoid). EMBO Molecular Medicine *3*: 334–347.

Cavill, R., Kamburov, A., Ellis, J.K., Athersuch, T.J., Blagrove, M.S.C., Herwig, R., et al. (2011). Consensus-phenotype integration of transcriptomic and metabolomic data implies a role for metabolism in the chemosensitivity of tumour cells. PLoS Comput Biol *7*: e1001113.

Chan, E.C., Koh, P.K., Mal, M., Cheah, P.Y., Eu, K.W., Backshall, A., et al. (2009). Metabolic Profiling of Human Colorectal Cancer Using High-Resolution Magic Angle Spinning Nuclear Magnetic Resonance (HR-MAS NMR) Spectroscopy and Gas Chromatography Mass Spectrometry (GC/MS). J. Proteome Res. *8*: 352–361.

Clevers, H., and Nusse, R. (2012). Wnt/β-Catenin Signaling and Disease. Cell *149*: 1192–1205.

Coulier, L., Bas, R., Jespersen, S., Verheij, E., van der Werf, M.J., and Hankemeier, T. (2006). Simultaneous Quantitative Analysis of Metabolites Using Ion-Pair Liquid Chromatography−Electrospray Ionization Mass Spectrometry. Anal. Chem. *78*: 6573–6582.

Cress, R.D., Morris, C., Ellison, G.L., and Goodman, M.T. (2006). Secular changes in colorectal cancer incidence by subsite, stage at diagnosis, and race/ethnicity, 1992–2001. Cancer *107*: 1142–1152.

Culhane, A.C., Perrière, G., Considine, E.C., Cotter, T.G., and Higgins, D.G. (2002). Between-group analysis of microarray data. Bioinformatics *18*: 1600–1608.

Culhane, A.C., Thioulouse, J., Perrière, G., and Higgins, D.G. (2005). MADE4: an R package for multivariate analysis of gene expression data. Bioinformatics *21*: 2789–2790.

Dai, M., Wang, P., Boyd, A.D., Kostov, G., Athey, B., Jones, E.G., et al. (2005). Evolving gene/transcript definitions significantly alter the interpretation of GeneChip data. Nucleic Acids Research *33*: e175–e175.

DeBerardinis, R.J., and Cheng, T. (2009). Q's next: the diverse functions of glutamine in metabolism, cell biology and cancer. Oncogene *29*: 313–324.

DeBerardinis, R.J., Sayed, N., Ditsworth, D., and Thompson, C.B. (2008). Brick by brick: metabolism and tumor cell growth. Curr Opin Genet Dev *18*: 54–61.

Denkert, C., Budczies, J., Weichert, W., Wohlgemuth, G., Scholz, M., Kind, T., et al. (2008). Metabolite profiling of human colon carcinoma – deregulation of TCA cycle and amino acid turnover. Mol Cancer *7*: 72.

Du, P., Kibbe, W.A., and Lin, S.M. (2006). Improved peak detection in mass spectrum by incorporating continuous wavelet transform-based pattern matching. Bioinformatics *22*: 2059–2065.

Dunn, W.B., Bailey, N.J., and Johnson, H.E. (2005). Measuring the metabolome: current analytical technologies. Analyst *130*: 606–625.

Ebara, T., Kaira, K., Saito, J.-I., Shioya, M., Asao, T., Takahashi, T., et al. (2010). L-type amino-acid transporter 1 expression predicts the response to preoperative hyperthermo-chemoradiotherapy for advanced rectal cancer. Anticancer Res. *30*: 4223–4227.

Fahy, E., Subramaniam, S., Murphy, R.C., Nishijima, M., Raetz, C.R.H., Shimizu, T., et al. (2009). Update of the LIPID MAPS comprehensive classification system for lipids. The Journal of Lipid Research *50 Suppl*: S9–14.

Farshidfar, F., Weljie, A.M., Kopciuk, K., Buie, W.D., MacLean, A., Dixon, E., et al. (2012). Serum metabolomic profile as a means to distinguish stage of colorectal cancer. Genome Med *4*: 42.

Fearon, E.R. (2011). Molecular genetics of colorectal cancer. Annu Rev Pathol *6*: 479–507.

Fearon, E.R., and Vogelstein, B. (1990). A genetic model for colorectal tumorigenesis. Cell *61*: 759–767.

Fiehn, O. (2002). Metabolomics - the link between genotypes and phenotypes. Plant Mol Biol *48*: 155–171.

Forcisi, S., Moritz, F., Kanawati, B., Tziotis, D., Lehmann, R., and Schmitt-Kopplin, P. (2013). Liquid chromatography-mass spectrometry in metabolomics research: mass analyzers in ultra high pressure liquid chromatography coupling. Journal of Chromatography A *1292*: 51–65.

Fuchs, B.C., and Bode, B.P. (2005). Amino acid transporters ASCT2 and LAT1 in cancer: partners in crime? Semin. Cancer Biol. *15*: 254–266.

Fuhrer, T., Heer, D., Begemann, B., and Zamboni, N. (2011). High-Throughput, Accurate Mass Metabolome Profiling of Cellular Extracts by Flow Injection–Time-of-Flight Mass Spectrometry. Anal. Chem. *83*: 7074–7080.

Gao, X., Zhang, Q., Da Meng, Isaac, G., Zhao, R., Fillmore, T.L., et al. (2012). A reversed-phase capillary ultra-performance liquid chromatography–mass spectrometry (UPLC-MS) method for comprehensive top-down/bottom-up lipid profiling. Anal Bioanal Chem *402*: 2923–2933.

Golub, T.R., Slonim, D.K., Tamayo, P., and Huard, C. (1999). Molecular classification of cancer: class discovery and class prediction by gene expression monitoring. Science *286*: 531–537.

Hanahan, D., and Weinberg, R.A. (2011). Hallmarks of Cancer: The Next Generation. Cell *144*: 646–674.

Hebert, A.S., Richards, A.L., Bailey, D.J., Ulbrich, A., Coughlin, E.E., Westphall, M.S., et al. (2014). The one hour yeast proteome. Molecular & Cellular Proteomics *13*: 339–347.

Hediger, M.A., Clémençon, B., Burrier, R.E., and Bruford, E.A. (2013). The ABCs of membrane transporters in health and disease (SLC series): introduction. Molecular Aspects of Medicine *34*: 95–107.

Hirayama, A., Kami, K., Sugimoto, M., Sugawara, M., Toki, N., Onozuka, H., et al. (2009). Quantitative Metabolome Profiling of Colon and Stomach Cancer Microenvironment by Capillary Electrophoresis Time-of-Flight Mass Spectrometry. Cancer Research *69*: 4918–4925.

Hollywood, K., Brison, D.R., and Goodacre, R. (2006). Metabolomics: Current technologies and future trends. Proteomics *6*: 4716–4723.

Hu, W., Zhang, C., Wu, R., Sun, Y., Levine, A., and Feng, Z. (2010). Glutaminase 2, a novel p53 target gene regulating energy metabolism and antioxidant function. Pnas *107*: 7455–7460.

Hummel, J., Segu, S., Li, Y., Irgang, S., Jueppner, J., and Giavalisco, P. (2011). Ultra Performance Liquid Chromatography and High Resolution Mass Spectrometry for the Analysis of Plant Lipids. Front. Plant Sci. *2*: 1–17.

Irizarry, R.A. (2003). Exploration, normalization, and summaries of high density oligonucleotide array probe level data. Biostatistics *4*: 249–264.

Ivanisevic, J., Zhu, Z.-J., Plate, L., Tautenhahn, R., Chen, S., O'Brien, P.J., et al. (2013). Toward 'Omic Scale Metabolite Profiling: A Dual Separation–Mass Spectrometry Approach for Coverage of Lipid and Central Carbon Metabolism. Anal. Chem. *85*: 130619135055004–6884.

Jain, M., Nilsson, R., Sharma, S., Madhusudhan, N., Kitami, T., Souza, A.L., et al. (2012). Metabolite Profiling Identifies a Key Role for Glycine in Rapid Cancer Cell Proliferation. Science *336*: 1040–1044.

Jemal, A., Bray, F., Center, Melissa M., Ferlay, J., Ward, E., and Forman, D. (2011). Global cancer statistics. CA Cancer J Clin *61*: 69–90.

Ji, H., Wu, G., Zhan, X., Nolan, A., Koh, C., De Marzo, A., et al. (2011). Cell-Type Independent MYC Target Genes Reveal a Primordial Signature Involved in Biomass Accumulation. PLoS ONE *6*: e26057.

Jones, S., Chen, W.-D., Parmigiani, G., Diehl, F., Beerenwinkel, N., Antal, T., et al. (2008). Comparative lesion sequencing provides insights into tumor evolution. Proc. Natl. Acad. Sci. U.S.a. *105*: 4283–4288.

Jozefczuk, S., Klie, S., Catchpole, G., Szymanski, J., Cuadros-Inostroza, Á., Steinhauser, D., et al. (2010). Metabolomic and transcriptomic stress response of Escherichia coli. Molecular Systems Biology *6*: 364.

Kamburov, A., Cavill, R., Ebbels, T.M.D., Herwig, R., and Keun, H.C. (2011). Integrated pathway-level analysis of transcriptomics and metabolomics data with IMPaLA. Bioinformatics *27*: 2917–2918.

Kandoth, C., McLellan, M.D., Vandin, F., Ye, K., Niu, B., Lu, C., et al. (2013). Mutational landscape and significance across 12 major cancer types. Nature *502*: 333–339.

Kenar, E., Franken, H., Forcisi, S., Wörmann, K., Häring, H.-U., Lehmann, R., et al. (2014). Automated label-free quantification of metabolites from liquid chromatography-mass spectrometry data. Molecular & Cellular Proteomics *13*: 348–359.

Kiefer, P., Delmotte, N., and Vorholt, J.A. (2010). Nanoscale Ion-Pair Reversed-Phase HPLC−MS for Sensitive Metabolome Analysis. Anal. Chem. *83*: 850–855.

Kitteringham, N.R., Jenkins, R.E., Lane, C.S., Elliott, V.L., and Park, B.K. (2009). Multiple reaction monitoring for quantitative biomarker analysis in proteomics and metabolomics. J. Chromatogr. B Analyt. Technol. Biomed. Life Sci. *877*: 1229–1239.

Kjeldahl, K., and Bro, R. (2010). Some common misunderstandings in chemometrics. J. Chemometrics *24*: 558–564.

Knittelfelder, O.L., Weberhofer, B.P., Eichmann, T.O., Kohlwein, S.D., and Rechberger, G.N. (2014). A versatile ultra-high performance LC-MS method for lipid profiling. J. Chromatogr. B Analyt. Technol. Biomed. Life Sci. *951-952*: 119–128.

Koek, M.M., Bakels, F., Engel, W., van den Maagdenberg, A., Ferrari, M.D., Coulier, L., et al. (2010). Metabolic profiling of ultrasmall sample volumes with GC/MS: from microliter to nanoliter samples. Anal. Chem. *82*: 156–162.

Koppenol, W.H., Bounds, P.L., and Dang, C.V. (2011). Otto Warburg's contributions to current concepts of cancer metabolism. Nature Reviews Cancer *11*: 325–337.

Kresnowati, M.T.A.P., van Winden, W.A., Almering, M.J.H., Pierick, ten, A., Ras, C., Knijnenburg, T.A., et al. (2006). When transcriptome meets metabolome: fast cellular responses of yeast to sudden relief of glucose limitation. Molecular Systems Biology *2*: 49.

Kroemer, G., and Pouyssegur, J. (2008). Tumor Cell Metabolism: Cancer‘s Achilles’ Heel. Cancer Cell *13*: 472–482.

Kuhl, C., Tautenhahn, R., Böttcher, C., Larson, T.R., and Neumann, S. (2012). CAMERA: an integrated strategy for compound spectra extraction and annotation of liquid chromatography/mass spectrometry data sets. Anal. Chem. *84*: 283–289.

Kunkel, M., Reichert, T.E., Benz, P., Lehr, H.A., Jeong, J.H., Wieand, S., et al. (2003). Overexpression of Glut-1 and increased glucose metabolism in tumors are associated with a poor prognosis in patients with oral squamous cell carcinoma. Cancer *97*: 1015–1024.

Lena, M.D. (2013). New strategies for colorectal cancer screening. World J. Gastroenterol. *19*: 1855–1860.

Li, B., and Simon, M.C. (2013). Molecular Pathways: Targeting MYC-induced metabolic reprogramming and oncogenic stress in cancer. Clin. Cancer Res. *19*: 5835–5841.

Li, S., Bin Guo, Song, J., Deng, X., Cong, Y., Li, P., et al. (2012). Plasma choline-containing phospholipids: potential biomarkers for colorectal cancer progression. Metabolomics *9*: 202–212.

Liu, W., Le, A., Hancock, C., Lane, A.N., Dang, C.V., Fan, T.W.M., et al. (2012). Reprogramming of proline and glutamine metabolism contributes to the proliferative and metabolic responses regulated by oncogenic transcription factor c-MYC. Proc. Natl. Acad. Sci. U.S.a. *109*: 8983–8988.

Locasale, J.W. (2013). Serine, glycine and one-carbon units: cancer metabolism in full circle. Nat Rev Cancer *13*: 572–583.

Locasale, J.W., Grassian, A.R., Melman, T., Lyssiotis, C.A., Mattaini, K.R., Bass, A.J., et al. (2011). Phosphoglycerate dehydrogenase diverts glycolytic flux and contributes to oncogenesis. Nat Genet *43*: 869–874.

Lu, W., Bennett, B.D., and Rabinowitz, J.D. (2008). Analytical strategies for LC–MS-based targeted metabolomics. Journal of Chromatography B-Analytical Technologies in the Biomedical and Life Sciences *871*: 236–242.

Luo, B., Groenke, K., Takors, R., Wandrey, C., and Oldiges, M. (2007). Simultaneous determination of multiple intracellular metabolites in glycolysis, pentose phosphate pathway and tricarboxylic acid cycle by liquid chromatography–mass spectrometry. Journal of Chromatography A *1147*: 153–164.

Lynch, H.T., and la Chapelle, de, A. (2003). Hereditary colorectal cancer. N. Engl. J. Med. *348*: 919–932.

Maglietta, R., Liuzzi, V.C., Cattaneo, E., Laczko, E., Piepoli, A., Panza, A., et al. (2012). Molecular pathways undergoing dramatic transcriptomic changes during tumor development in the human colon. BMC Cancer *12*: 608.

Mann, M., Hendrickson, R.C., and Pandey, A. (2001). Analysis of proteins and proteomes by mass spectrometry. Annu. Rev. Biochem. *70*: 437–473.

Manna, S.K., Tanaka, N., Krausz, K.W., Haznadar, M., Xue, X., Matsubara, T., et al. (2014). Biomarkers of Coordinate Metabolic Reprogramming in Colorectal Tumors in Mice and Humans. Gastroenterology.

Marra, G., and Schar, P. (1999). Recognition of DNA alterations by the mismatch repair system. Biochem. J. *338*: 1–13.

Meijer, T.W.H., Kaanders, J.H.A.M., Span, P.N., and Bussink, J. (2012). Targeting Hypoxia, HIF-1, and Tumor Glucose Metabolism to Improve Radiotherapy Efficacy. Clincancerres.Aacrjournals.org *18*: 5585–5594.

Milne, S.B., Mathews, T.P., Myers, D.S., Ivanova, P.T., and Brown, H.A. (2013). Sum of the Parts: Mass Spectrometry-Based Metabolomics. Biochemistry *52*: 3829–3840.

Muzny, D.M., Bainbridge, M.N., Chang, K., Dinh, H.H., Drummond, J.A., Fowler, G., et al. (2012). Comprehensive molecular characterization of human colon and rectal cancer. Nature *487*: 330–337.

Myint, K.T., Aoshima, K., Tanaka, S., Nakamura, T., and Oda, Y. (2009a). Quantitative Profiling of Polar Cationic Metabolites in Human Cerebrospinal Fluid by Reversed-Phase Nanoliquid Chromatography/Mass Spectrometry. Anal. Chem. *81*: 1121–1129.

Myint, K.T., Uehara, T., Aoshima, K., and Oda, Y. (2009b). Polar Anionic Metabolome Analysis by Nano-LC/MS with a Metal Chelating Agent. Anal. Chem. *81*: 7766–7772.

Nomura, D.K., Long, J.Z., Niessen, S., Hoover, H.S., Ng, S.W., and Cravatt, B.F. (2010). Monoacylglycerol Lipase Regulates a Fatty Acid Network that Promotes Cancer Pathogenesis. Cell *140*: 49–61.

Nordström, A., O'Maille, G., Qin, C., and Siuzdak, G. (2006). Nonlinear data alignment for UPLC-MS and HPLC-MS based metabolomics: quantitative analysis of endogenous and exogenous metabolites in human serum. Anal. Chem. *78*: 3289–3295.

Ogata, H., Goto, S., Sato, K., Fujibuchi, W., Bono, H., and Kanehisa, M. (1999). KEGG: Kyoto Encyclopedia of Genes and Genomes. Nucleic Acids Research *27*: 29–34.

Ong, E.S., Zou, L., Li, S., Cheah, P.Y., Eu, K.W., and Ong, C.N. (2010). Metabolic profiling in colorectal cancer reveals signature metabolic shifts during tumorigenesis. Molecular & Cellular Proteomics.

Oshima, M., Dinchuk, J.E., Kargman, S.L., Oshima, H., Hancock, B., Kwong, E., et al. (1996). Suppression of intestinal polyposis in Apc delta716 knockout mice by inhibition of cyclooxygenase 2 (COX-2). Cell *87*: 803–809.

Parsons, D.W., Jones, S., Zhang, X., Lin, J.C.-H., Leary, R.J., Angenendt, P., et al. (2008). An Integrated Genomic Analysis of Human Glioblastoma Multiforme. Science *321*: 1807–1812.

Participants in the Paris Workshop (2003). The Paris endoscopic classification of superficial neoplastic lesions: esophagus, stomach, and colon. Gastrointestinal Endoscopy *58*: S3–S43.

Patti, G.J., Tautenhahn, R., Rinehart, D., Cho, K., Shriver, L.P., Manchester, M., et al. (2013). A view from above: cloud plots to visualize global metabolomic data. Anal. Chem. *85*: 798–804.

Patti, G.J., Yanes, O., and Siuzdak, G. (2012). Innovation: Metabolomics: the apogee of the omics trilogy. Nat. Rev. Mol. Cell Biol. *13*: 263–269.

Pesek, J.J., Matyska, M.T., Fischer, S.M., and Sana, T.R. (2008). Analysis of hydrophilic metabolites by high-performance liquid chromatography–mass spectrometry using a silica hydride-based stationary phase. Journal of Chromatography A *1204*: 48–55.

Phang, J.M., Liu, W., Hancock, C., and Christian, K.J. (2012). The proline regulatory axis and cancer. Front Oncol *2*: 60.

Plumb, R., Castro-Perez, J., Granger, J., Beattie, I., Joncour, K., and Wright, A. (2004). Ultra-performance liquid chromato... [Rapid Commun Mass Spectrom. 2004] - PubMed - NCBI. Rapid Commun. Mass Spectrom. *18*: 2331–2337.

Plumb, R.S., Johnson, K.A., Rainville, P., Smith, B.W., Wilson, I.D., Castro-Perez, J.M., et al. (2006). UPLC/MSE; a new approach for generating molecular fragment information for biomarker structure elucidation. Rapid Commun. Mass Spectrom. *20*: 1989–1994.

Possemato, R., Marks, K.M., Shaul, Y.D., Pacold, M.E., Kim, D., Birsoy, K., et al. (2011). Functional genomics reveal that the serine synthesis pathway is essential in breast cancer. Nature *476*: 346–350.

Prince, J.T., and Marcotte, E.M. (2006). Chromatographic Alignment of ESI-LC-MS Proteomics Data Sets by Ordered Bijective Interpolated Warping. Anal. Chem. *78*: 6140–6152.

R Development Core Team (2013). R: A language and environment for statistical computing (Vienna, Austria).

Richard T Gallagher, Michael P Balogh, Paul Davey, Mike R Jackson, Ian Sinclair, A., and Southern, L.J. (2003). Combined Electrospray Ionization−Atmospheric Pressure Chemical Ionization Source for Use in High-Throughput LC−MS Applications. Anal. Chem. *75*: 973–977.

Roepenack-Lahaye, von, E., Degenkolb, T., Zerjeski, M., Franz, M., Roth, U., Wessjohann, L., et al. (2004). Profiling of Arabidopsis secondary metabolites by capillary liquid chromatography coupled to electrospray ionization quadrupole time-of-flight mass spectrometry. Plant Physiol. *134*: 548–559.

Römisch-Margl, W., Prehn, C., Bogumil, R., Röhring, C., Suhre, K., and Adamski, J. (2011). Procedure for tissue sample preparation and metabolite extraction for high-throughput targeted metabolomics. Metabolomics *8*: 133–142.

Sabates Bellver, J., Van der Flier, L.G., de Palo, M., Cattaneo, E., Maake, C., Rehrauer, H., et al. (2007). Transcriptome profile of human colorectal adenomas. Mol. Cancer Res. *5*: 1263–1275.

Sansom, O.J., Meniel, V.S., Muncan, V., Phesse, T.J., Wilkins, J.A., Reed, K.R., et al. (2007). Myc deletion rescues Apc deficiency in the small intestine. Nature *446*: 676–679.

Shannon, P., Markiel, A., Ozier, O., Baliga, N.S., Wang, J.T., Ramage, D., et al. (2003). Cytoscape: a software environment for integrated models of biomolecular interaction networks. Genome Res. *13*: 2498–2504.

Shen, Y., Zhao, R., Berger, S.J., Anderson, G.A., Rodriguez, N., and Smith, R.D. (2002). High-efficiency nanoscale liquid chromatography coupled on-line with mass spectrometry using nanoelectrospray ionization for proteomics. Anal. Chem. *74*: 4235–4249.

Shim, B.Y., Jung, J.-H., Lee, K.-M., Kim, H.-J., Hong, S.H., Kim, S.H., et al. (2013). Glucose transporter 1 (GLUT1) of anaerobic glycolysis as predictive and prognostic values in neoadjuvant chemoradiotherapy and laparoscopic surgery for locally advanced rectal cancer. International Journal of Colorectal Disease *28*: 375–383.

Siegel, R., Naishadham, D., and Jemal, A. (2013). Cancer statistics, 2013. CA Cancer J Clin *63*: 11–30.

Smith, C.A., Want, E.J., O'Maille, G., Abagyan, R., and Siuzdak, G. (2006). XCMS: Processing Mass Spectrometry Data for Metabolite Profiling Using Nonlinear Peak Alignment, Matching, and Identification. Anal. Chem. *78*: 779–787.

Smith, R., Ventura, D., and Prince, J.T. (2013). LC-MS alignment in theory and practice: a comprehensive algorithmic review. Brief. Bioinformatics.

Snyder, L.R., Kirkland, J.J., and Dolan, J.W. (2011). Introduction to Modern Liquid Chromatography (John Wiley & Sons).

Som, P., Atkins, H.L., Bandoypadhyay, D., Fowler, J.S., MacGregor, R.R., Matsui, K., et al. (1980). A fluorinated glucose analog, 2-fluoro-2-deoxy-D-glucose (F-18): nontoxic tracer for rapid tumor detection. J Nucl Med *21*: 670–675.

Son, J., Lyssiotis, C.A., Ying, H., Wang, X., Hua, S., Ligorio, M., et al. (2013). Glutamine supports pancreatic cancer growth through a KRAS-regulated metabolic pathway. Nature *496*: 101–105.

Stratton, M.R., Campbell, P.J., and Futreal, P.A. (2009). The cancer genome. Nature *458*: 719–724.

Sugimoto, M., Kawakami, M., Robert, M., Soga, T., and Tomita, M. (2012). Bioinformatics Tools for Mass Spectroscopy-Based Metabolomic Data Processing and Analysis. Curr Bioinform *7*: 96–108.

Suzuki, S., Tanaka, T., Poyurovsky, M.V., Nagano, H., Mayama, T., Ohkubo, S., et al. (2010). Phosphate-activated glutaminase (GLS2), a p53-inducible regulator of glutamine metabolism and reactive oxygen species. Pnas *107*: 7461–7466.

Szablewski, L. (2013). Expression of glucose transporters in cancers. Biochimica Et Biophysica Acta (BBA) - Reviews on Cancer *1835*: 164–169.

Tautenhahn, R. (2009). Feature-Detektion, Annotation und Alignment von Metabolomik LC-MS-Daten.

Tautenhahn, R., Böttcher, C., and Neumann, S. (2008). Highly sensitive feature detection for high resolution LC/MS. BMC Bioinformatics *9*: 504.

Tautenhahn, R., Cho, K., Uritboonthai, W., Zhu, Z., Patti, G.J., and Siuzdak, G. (2012). An accelerated workflow for untargeted metabolomics using the METLIN database. Nat. Biotechnol. *30*: 826–828.

Thioulouse, J., CHESSEL, D., DEC, S.D., and OLIVIER, J.-M. (1997). ADE-4: a multivariate analysis and graphical display software - Springer. Stat Comput *7*: 75–83.

Tomlinson, I.P.M., Alam, N.A., Rowan, A.J., Barclay, E., Jaeger, E.E.M., Kelsell, D., et al. (2002). Germline mutations in FH predispose to dominantly inherited uterine fibroids, skin leiomyomata and papillary renal cell cancer. Nat Genet *30*: 406–410.

Truninger, K., Menigatti, M., Luz, J., Russell, A., and Haider, R. (2005). Immunohistochemical analysis reveals high frequency of PMS2 defects in colorectal cancer. Gastroenterology *128*: 1160–1171.

Vander Heiden, M.G., Cantley, L.C., and Thompson, C.B. (2009). Understanding the Warburg Effect: The Metabolic Requirements of Cell Proliferation. Science *324*: 1029–1033.

Vogelstein, B., and Kinzler, K.W. (2004). Cancer genes and the pathways they control. Nat Med *10*: 789–799.

Vogelstein, B., Papadopoulos, N., Velculescu, V.E., Zhou, S., Diaz, L.A., and Kinzler, K.W. (2013). Cancer genome landscapes. Science *339*: 1546–1558.

Warburg, O. (1925). The metabolism of carcinoma cells. The Journal of Cancer Research *9*: 148–163.

Warburg, O. (1956). On the origin of cancer cells. Science *123*: 309–314.

Warburg, O.H. (1926). Über den Stoffwechsel der Tumoren (Julius Springer Verlag, Berlin).

Ward, P.S., and Thompson, C.B. (2012). Metabolic Reprogramming: A Cancer Hallmark Even Warburg Did Not Anticipate. Cancer Cell *21*: 297–308.

Wei, X., Sun, W., Shi, X., Koo, I., Wang, B., Zhang, J., et al. (2011). MetSign: A Computational Platform for High-Resolution Mass Spectrometry-Based Metabolomics. Anal. Chem. *83*: 7668–7675.

Weinberg, R.A. (1994). Oncogenes and tumor suppressor genes. CA Cancer J Clin *44*: 160–170.

Weinhouse, S. (1976). The Warburg hypothesis fifty years later. Z. Krebsforsch. *87*: 115–126.

Weinhouse, S., Warburg, O., Burk, D., and Schade, A.L. (1956). On respiratory impairment in cancer cells. Science *124*: 267–271.

Wilkins, J.A., and Sansom, O.J. (2008). C-Myc is a critical mediator of the phenotypes of Apc loss in the intestine. Cancer Research *68*: 4963–4966.

Winawer, S.J., Zauber, A.G., Ho, M.N., O'Brien, M.J., Gottlieb, L.S., Sternberg, S.S., et al. (1993). Prevention of colorectal cancer by colonoscopic polypectomy. The National Polyp Study Workgroup. N. Engl. J. Med. *329*: 1977–1981.

Wise, D.R., DeBerardinis, R.J., Mancuso, A., Sayed, N., Zhang, X.Y., Pfeiffer, H.K., et al. (2008). Myc regulates a transcriptional program that stimulates mitochondrial glutaminolysis and leads to glutamine addiction. Proc. Natl. Acad. Sci. U.S.a. *105*: 18782–18787.

Wishart, D.S., Jewison, T., Guo, A.C., Wilson, M., Knox, C., Liu, Y., et al. (2013). HMDB 3.0--The Human Metabolome Database in 2013. Nucleic Acids Research *41*: D801–7.

Wishart, D.S., Tzur, D., Knox, C., Eisner, R., Guo, A.C., Young, N., et al. (2007). HMDB: the Human Metabolome Database. Nucleic Acids Research *35*: D521–6.

Yanes, O., Tautenhahn, R., Patti, G.J., and Siuzdak, G. (2011). Expanding Coverage of the Metabolome for Global Metabolite Profiling. Anal. Chem. *83*: 2152–2161.

Ying, H., Kimmelman, A.C., Lyssiotis, C.A., Hua, S., Chu, G.C., Fletcher-Sananikone, E., et al. (2012). Oncogenic Kras maintains pancreatic tumors through regulation of anabolic glucose metabolism. Cell *149*: 656–670.

Abbreviations

ADP	Adenosine diphosphate ribose
APC	Adenomatous polyposis coli
ATP	Adenosintriphosphat
BEH	Ethylene Bridged Hybrid
BGA	Between group analysis
BPI	Base peak intensity
capLCMS	Capillary liquid chromatography mass spectrometry
CRC	Colorectal cancer
CWT	continuous wavelet transformation
DDA	data dependent acquisition
DG	Diradylglycerols
DIA	data independent acquisition
DNA	Deoxyribonucleic acid
EIC	Extracted ion chromatogram
ER	Endoplasmic reticulum
ESI	Electrospray ionization
FAP	Familial adenomatous polyposis
FASN	Fatty acid synthase
FWHM	Full width at half maximum
GPI	Glycerophosphoinositol
HILIC	Hydrophilic interaction chromatography
HMDB	Human metabolome database
HNPCC	Hereditary nonpolyposis colorectal cancer
HPLC	High-performance liquid chromatography
HSS	High Strength Silica
KEGG	Kyoto Encyclopedia of Genes and Genomes
KGML	KEGG markup language
LC	Liquid chromatography
LCMS	Liquid chromatography mass spectrometry
LOD	Limit of detection
logP	Partition coefficient
m/z	mass-to-charge ratio
MeOH	Methanol
MRM	Multiple reaction monitoring
MS	Mass spectrometry
MS/MS	Tandem mass spectrometry
MSE	Waters brand name for DIA (MS plus "energy")
MSI	Microsatellite instability
nanoESI	Nano-electrospray ionization mass spectrometry
nanoLCMS	Nanoscale liquid chromatography coupled to tandem mass spectrometry
NH4	Ammonium
NH4Ac	Ammonium acetate
NMR	Nuclear magnetic resonance
PCA	Principal component analysis

PRPP	Phosphoribosyl pyrophosphate
RMA	Robust Multi-array Average
RNA	Ribonucleic acid
RP	Reversed phase
RSD	Relative standard deviation
RT	Retention time
S/N	Signal to noise
SNR	Signal to noise ratio
TBA	Tributylamine
TG	Triradylglycerol
TOF	Time of flight
UHPLC/UPLC	Ultra Performance Liquid Chromatography
VLC-PUFA	Very-long-chain polyunsaturated fatty acid

Acknowledgements

First, I would like to thank Prof. Josef Jiricny that he accepted to take the responsibility for me as a PhD student and for the scientific support he gave me during the lab meetings.

I say thank you to both of my supervisors Dr. Endre Laczko and PD Dr. Giancarlo Marra. Both gave me to opportunity to work on an exciting project and enabled me to independently find my way through the world of cancer metabolomics.

I further thank my committee members, Prof. Uwe Sauer and Prof. Vassily Hatzimanikatis for valuable feedback.

Thank you to Prof. Reinhard Furrer who always gave me help when I needed assistance in data analysis. He also kindly agreed to act both as committee member and referee for this dissertation.

I further want to acknowledge the expert review efforts by Dr. Christian Frezza, who agreed to function as an external referee.

Thank you to Elisa Cattaneo and the collaboration partners in Bari, Cremona and St. Giovanni Rotondo for making the transcriptome data available.

Thank you to Prof. Peter Bauerfeind and Dr. Christine Manser for the collaboration with the collection of biopsy samples.

A big thanks goes to all members of the FGCZ, which became home to me for the last 5 years. Especially I want to thank Prof. Ralph Schlapbach for supporting me as full member of his institute. The 11 o'clock lunch team and the "can't score" team I thank for a great time.

The finalization of this thesis needed an extra portion of endurance, and I am very thankful to Sabrina. She gave me all her love and support to keep me on track.

Curriculum vitae

David Fischer, Diplombiologe
Born December 1st, 1983 in Villingen-Schwenningen, Germany

Winzerstrasse 7 CH-8049 Zurich +41 76 584 92 53 David.Fischer@uzh.ch

Education

Apr 2009 – Mar 2014	**PhD student, Functional Genomics Center, University of Zurich** Development of new analytical and bioinformatical techniques for mass spectrometry based metabolomics. Systems biology of colorectal cancer. Member of the Cancer Biology PhD Program in Zurich. Supervision: Dr. Endre Laczko
Oct 2008	**Diploma in Biology, University of Konstanz** Characterization of the enzyme acetylene hydratase in aerobic and anaerobic bacteria. Grade 1.0 (highest grade). Supervision: Prof. Dr. Peter Kroneck
Oct 2005 – Oct 2008	**Advanced studies in Biology, University of Konstanz** Study focus: Organic Biochemistry, Biomathematics, Structural Biology, Cellular Biochemistry.
Oct 2003 – Oct 2005	**Basic studies in Biology (Vordiplom), University of Konstanz**
Jun 2003	**Abitur, Albertus Magnus Gymnasium Rottweil, Germany** Grade "good"
Sep 1994 – Jun 2003	**High school, Albertus Magnus Gymnasium Rottweil, Germany**
Sep 1990 – Sep 1994	**Primary School, Deisslingen, Germany**

Other professional activities

Feb 2007 – Apr 2007	**Internship position, Max Planck Institute for Biophysics, Frankfurt** Structural analysis of the c-ring membrane protein of a cyanobacterial F_0F_1-ATP synthase.
July 2006 – Aug 2006	**Volunteer position, Bilsa Biological Station, Ecuador** Reforestation work in primary cloud forests.
Jun 2005 – Jan 2008	**Student assistant position in environmental physics, Institute for Limnology, University of Konstanz** Software development for the evaluation of surface wave data.
Oct 2004 – Jun 2006	**Student assistant position as lecturer** Undergraduate student courses in biomathematics
Computer skills	MATLAB (10 years), R (5 years), Microsoft office
Languages	German (native), English (fluent)

Publications

F. Rosenthal, K. L. H. Feijs, E. Frugier, M. Bonalli, A. H. Forst, R. Imhof, H. C. Winkler, D. Fischer, A. Caflisch, P. O. Hassa, B. Lüscher, and M. O. Hottiger, "Macrodomain-containing proteins are new mono-ADP-ribosylhydrolases.," *Nat. Struct. Mol. Biol.*, vol. 20, no. 4, pp. 502–507, Apr. 2013.

J.-H. Jang, A. Rickenbacher, B. Humar, A. Weber, D. A. Raptis, K. Lehmann, B. Stieger, W. Moritz, C. Soll, P. Georgiev, D. Fischer, E. Laczko, R. Graf, and P.-A. Clavien, "Serotonin protects mouse liver from cholestatic injury by decreasing bile salt pool after bile duct ligation.," *Hepatology*, vol. 56, no. 1, pp. 209–218, Jul. 2012.

G. B. Seiffert, D. Abt, F. tenBrink, D. Fischer, O. Einsle, and P. M. Kroneck, "Acetylene Hydratase," in *Handbook of Metalloproteins*, Chichester, UK: John Wiley & Sons, Ltd, 2006.

Publications in preparation

E. Laczko, J. Hu, S. Hartnack, D. Fischer, B. Riond, C. Reusch, M. Ackermann, "Alterations of Serum Free Fatty Acid and Phospholipid Levels in Feline Diabetes using Ultra Performance Liquid Chromatography coupled with TOF Mass Spectrometry", submitted to PLOS ONE

D. Fischer, G. Marra and E. Laczko, "Capillary ultra-performance liquid chromatography-mass spectrometry for fast and sensitive metabolome analysis", to be submitted

D. Fischer, C. Panse and E. Laczko, "Improving the detection of low abundant metabolites by combining ion intensities of multiple LC/MS runs", to be submitted

D. Fischer, C. Manser, P. Bauerfeind, G. Marra and E. Laczko, "Metabolic alterations during colorectal cancer development – a combined analysis of the metabolome and transcriptome", to be submitted

Zurich, March 19, 2014